**Conocimiento de las fuentes de datos:**

¿Cuántas fuentes de datos existen?

Además de **la fuente de conocimiento que constituyen los distintos miembros del proyecto** (los cuáles se listan a continuación), gracias a su conocimiento, y a la posibilidad de transmitirlo para el avance del proyecto:

-BI Devs, con conocimientos en ETL, limpieza de datos, preprocesamiento de datos, modelamiento de datos, aprendizaje de máquina: supervisado, no supervisado, y profundo.

-BI Analyst, con conocimientos en estadística básica (medidas de tendencia central, de dispersión, distribuciones, pruebas de hipótesis, correlaciones), exploración visual de datos, evaluación y selección de modelos de datos.

-Data Scientist, con conocimientos avanzados en infraestructura y despliegue de servicios para consumo y almacenamiento de datos (basados en AWS), despliegue de APIs REST, despliegue de servicios de gestores de bases de datos como PostgreSQL y MongoDB. Recurrir a él para soporte técnico de infraestructura.

-Product Owner, y Scrum Master, conocedores de la metodología Scrum y familiarizados con las tareas técnicas a desarrollar en el proyecto; para definirlas y vigilar su ejecución. El product owner, naturalmente, también está al tanto de los requerimientos comerciales y de patrocinadores; y a partir de ello define las actividades del proyecto.

**Tenemos también como fuentes de datos para el proyecto:**

Un repositorio con archivos de texto con datos no estructurados obtenidos mediante webscrapping y una etapa inicial de procesamiento; la cual consistió en remover las etiquetas HTML y el contenido ajeno a la información divulgativa dispuesta para el usuario final del sitio web de donde fueron extraídos. Esta etapa inicial de procesamiento, sin embargo, no limpió completamente el resultado de caracteres especiales, y símbolos ajenos a las palabras del idioma español.

El contenido de los archivos del repositorio corresponde al texto de artículos de divulgación académica que fueron originalmente publicados por terceros en el portal Wikipedia. Los artículos de Wikipedia fueron seleccionados aleatoriamente; por tanto, no guardan relación.

Estos archivos fueron provistos por el cliente para que en el proyecto se pueda extraer un banco de palabras de acuerdo a los siguientes lineamientos:

1. No incluir palabras con más de una letra en mayúscula (acrónimos).

2. No incluir stop words.

3. No incluir palabras con menos de 4 caracteres.

2. No permitir letras con tildes. Reemplazar letras con tildes por sus homólogos sin ellas.

3. No permitir caracteres de formato especiales (ej: "\t", "\n", "\u200b"), símbolos (@, !, ?, etc...), y letras ajenas al lenguaje español. Reemplazarlos por espacios en blanco.

8. Finalmente todas las palabras obtenidas deben estar en minúscula.

El análisis del banco de palabras a construir con los archivos de la fuente de datos indicada, y su procesamiento, constituirá el conjunto de datos de referencia para los sistemas predictores de palabras (las palabras fuera de este banco no podrán ser predichas por los predictores a desarrollar), y será lo que fundamentará los lineamientos básicos de su lógica de predicción.

Como puede entenderse a partir de lo explicado anteriormente, los datos de la fuente descrita no se encuentran en un formato estructurado, deduplicado, ni conforme a los requerimientos del cliente. Por otro lado, el proceso de limpieza podría comprometer su completitud. Estas condiciones perjudican su calidad.

Adicionalmente, **para futuras etapas del proyecto se contará con las siguientes fuentes de datos**:

Base de datos tipo PostgreSQL alojada en Render; la cual contendrá los resultados de las pruebas de cada método predictor de palabras implementado, así como un directorio de datos con las palabras del banco de referencia asociadas a su id. Las tablas con los resultados de las pruebas deberán poblarse a medida que se realicen pruebas con un predictor, siguiendo los lineamientos definidos para almacenar datos de calidad en la misma.

Base de datos no SQL en MongoDB; la cual contendrá información sobre cada usuario de método predictor de palabras, y los intentos de predicción realizados por cada usuario; de acuerdo con una estructura definida previamente. Los documentos de esta base de datos se poblarán automáticamente a medida que un usuario consumas las API del sistema Wordle, siguiendo los lineamientos definidos por el cliente para almacenar datos de calidad en la misma. Una de las particularidades de esta base de datos, es que sus documentos registrarán las fechas en que un usuario consumió las Api del sistema Wordle; esto permitirá, más adelante, medir la agilidad con la que los usuarios logran predecir palabras.

Aunque estas bases de datos se actualizan de manera independiente, las tablas o documentos que almacenarán información sobre los intentos de predicción en cada una de ellas deberán compartir la hora en que se realizó cada intento, el id de la palabra que se intentaba predecir, y la palabra estimada en la predicción.

**Recolección inicial de datos:**

Tal como se mencionaba anteriormente, los archivos dispuestos por el cliente para la búsqueda de palabras pueden encontrarse en un repositorio público en Github, mediante el siguiente enlace:

https://github.com/amendivilsejin/training\_analitica\_2nd\_module\_repo/tree/master/raw\_texts

En total, fueron provistos 65 archivos de texto con textos para divulgación académica que son ilustrativos del conjunto de vocablos que forman la lengua española; por lo que se consideran suficientes para los posteriores procesos de análisis del proyecto. En total, estos archivos pesan 2,65 MB. Los archivos fueron obtenidos mediante webscrapping y una etapa inicial de procesamiento; la cual consistió en remover las etiquetas HTML y el contenido ajeno a la información divulgativa dispuesta para el usuario final del sitio web de donde fueron extraídos. Esta etapa inicial de procesamiento, sin embargo, no limpió completamente el resultado de caracteres especiales, y símbolos ajenos a las palabras del idioma español.

Para poder utilizar cada uno de estos archivos, basta con utilizar un programa que sea capaz de utilizar las rutas o nombres de cada uno para leerlos; esto, siempre que el acceso a los mismos no esté restringido. Por el presente proyecto, se realizará un script de Python que utilice las funcionalidades nativas de este lenguaje de programación (como la función open) para recorrer la lista de los nombres de los archivos y leerlos uno a uno.

**Muestreo de datos:**

Para realizar un muestreo de todos los archivos de la fuente de datos que podemos revisar en este punto del proyecto (repositorio con archivos de texto), primero definimos las rutas de los distintos archivos a leer:

files\_address\_base = "/content/drive/MyDrive/Colab Notebooks/data/raw\_texts/"

source\_files\_addresses = [

  files\_address\_base + "\_C3\_81frica.txt",

  files\_address\_base + "Aeropuerto.txt",

  files\_address\_base + "Agricultura.txt",

  files\_address\_base + "Alan\_Turing.txt",

  files\_address\_base + "Antiguo\_Egipto.txt",

  files\_address\_base + "Arte.txt",

  files\_address\_base + "Asia.txt",

  files\_address\_base + "Babilonia\_(ciudad).txt",

  files\_address\_base + "Biblia.txt",

  files\_address\_base + "Biodiversidad.txt",

  files\_address\_base + "Capitalismo.txt",

  files\_address\_base + "Ciencia.txt",

  files\_address\_base + "Colombia.txt",

  files\_address\_base + "Consumo.txt",

  files\_address\_base + "Cristianismo.txt",

  files\_address\_base + "Cristo.txt",

  files\_address\_base + "Cultura.txt",

  files\_address\_base + "Democracia.txt",

  files\_address\_base + "Diccionario.txt",

  files\_address\_base + "Divina\_comedia.txt",

  files\_address\_base + "Ecologismo.txt",

  files\_address\_base + "Energ\_C3\_ADa\_nuclear.txt",

  files\_address\_base + "Enrique\_VIII\_de\_Inglaterra.txt",

  files\_address\_base + "Escritura.txt",

  files\_address\_base + "Est\_C3\_A9tica.txt",

  files\_address\_base + "Estado.txt",

  files\_address\_base + "Estados\_Unidos.txt",

  files\_address\_base + "Europa.txt",

  files\_address\_base + "Filosof\_C3\_ADa.txt",

  files\_address\_base + "Finanzas.txt",

  files\_address\_base + "Fruta.txt",

  files\_address\_base + "Gabriel\_Garc\_C3\_ADa\_M\_C3\_A1rquez.txt",

  files\_address\_base + "Gastronom\_C3\_ADa\_de\_M\_C3\_A9xico.txt",

  files\_address\_base + "George\_Patton.txt",

  files\_address\_base + "Grecia.txt",

  files\_address\_base + "Guerra\_Fr\_C3\_ADa.txt",

  files\_address\_base + "Henry\_Ford.txt",

  files\_address\_base + "Holocausto.txt",

  files\_address\_base + "Homo\_sapiens.txt",

  files\_address\_base + "Inteligencia\_artificial.txt",

  files\_address\_base + "Italia.txt",

  files\_address\_base + "Jorge\_Isaacs.txt",

  files\_address\_base + "Juda\_C3\_ADsmo.txt",

  files\_address\_base + "Juglar.txt",

  files\_address\_base + "Julio\_Garavito.txt",

  files\_address\_base + "Liberalismo.txt",

  files\_address\_base + "Libro.txt",

  files\_address\_base + "Ludwig\_van\_Beethoven.txt",

  files\_address\_base + "Nevado\_del\_Ruiz.txt",

  files\_address\_base + "Nikola\_Tesla.txt",

  files\_address\_base + "Odisea.txt",

  files\_address\_base + "Organizaci\_C3\_B3n\_de\_las\_Naciones\_Unidas.txt",

  files\_address\_base + "Per\_C3\_ADodo\_helen\_C3\_ADstico.txt",

  files\_address\_base + "Pizza.txt",

  files\_address\_base + "Pueblos\_germ\_C3\_A1nicos.txt",

  files\_address\_base + "Real\_Academia\_Espa\_C3\_B1ola.txt",

  files\_address\_base + "Reino\_Unido.txt",

  files\_address\_base + "Revoluci\_C3\_B3n\_Industrial.txt",

  files\_address\_base + "Rueda.txt",

  files\_address\_base + "Salsa\_(g\_C3\_A9nero\_musical).txt",

  files\_address\_base + "Saxof\_C3\_B3n.txt",

  files\_address\_base + "Segunda\_Guerra\_Mundial.txt",

  files\_address\_base + "Sublime.txt",

  files\_address\_base + "Thomas\_Alva\_Edison.txt",

  files\_address\_base + "Valle.txt"

]

Una vez realizado esto, procedemos a realizar el muestreo de cada archivo mediante el siguiente ciclo:

for address in source\_files\_addresses:

  print("Sample from " + address[len(files\_address\_base):] + ":\n")

  sample\_txt\_file(address)

  print("-----------------------------------------------------------------")

A continuación, presentamos una lista con el muestreo de los primeros 200 caracteres de cada fuente de datos:

Sample from \_C3\_81frica.txt:

Reunión (Francia)

Mayotte (Francia)

Santa Elena, Ascensión y Tristán de Acuña

Madeira

Pantelaria

Islas Canarias

Ceuta

Melilla

África es el tercer continente más extenso, tras Asia y América. Est

-----------------------------------------------------------------

Sample from Aeropuerto.txt:

Un aeropuerto es un aeródromo con terminal para el viaje de transporte aéreo en aeronave. Las funciones del aeropuerto son variadas, entre ellas el aterrizaje y despegue de aeronaves, embarque y desem

-----------------------------------------------------------------

Sample from Agricultura.txt:

La agricultura (del latín agri ‘campo’ y cultūra ‘cultivo’, ‘crianza’)[1]​[2]​ es el conjunto de actividades económicas y técnicas relacionadas con el tratamiento del suelo y el cultivo de la tierra p

-----------------------------------------------------------------

Sample from Alan\_Turing.txt:

Alan Mathison Turing (Paddington, Londres; 23 de junio de 1912-Wilmslow, Cheshire; 7 de junio de 1954) fue un matemático, lógico, informático teórico, criptógrafo, filósofo y biólogo teórico británico

-----------------------------------------------------------------

Sample from Antiguo\_Egipto.txt:

El Antiguo Egipto o Egipto Antiguo fue una civilización de la Antigüedad, que se originó a lo largo del cauce medio y bajo del río Nilo, cuya historia abarca más de tres milenios. Se la considera una

-----------------------------------------------------------------

Sample from Arte.txt:

El arte (del latín ars, artis, y este calco del griego τέχνη téchnē)[1]​ es entendido generalmente como cualquier actividad o producto realizado con una finalidad estética y también comunicativa, me

-----------------------------------------------------------------

Sample from Asia.txt:

Países transcontinentales: Armenia Azerbaiyán Chipre Egipto Georgia Indonesia Kazajistán Rusia Turquía

Asia es el continente más grande y poblado de la Tierra. Con cerca de 45 millones de km², supone

-----------------------------------------------------------------

Sample from Babilonia\_(ciudad).txt:

Babilonia fue una antigua ciudad de la Baja Mesopotamia situada cerca de la actual ciudad de Hilla (Irak). Fue la capital del antiguo Reino babilónico, y por varios siglos fue considerada como un impo

-----------------------------------------------------------------

Sample from Biblia.txt:

La Biblia (del latín biblĭa, y este del griego βιβλία biblía, ‘libros’)[1]​ es un conjunto de libros canónicos que en el cristianismo y en otras religiones, se consideran producto de inspiración divin

-----------------------------------------------------------------

Sample from Biodiversidad.txt:

La biodiversidad o diversidad biológica es, según el Convenio Internacional sobre la Diversidad Biológica, el término por el que se hace referencia a la amplia variedad de seres vivos sobre la Tierra

-----------------------------------------------------------------

Sample from Capitalismo.txt:

El capitalismo es un orden o sistema social y económico que se encuentra en constante movimiento, derivado del usufructo de la propiedad privada sobre el capital como herramienta de producción, que se

-----------------------------------------------------------------

Sample from Ciencia.txt:

La ciencia (del latín scientĭa, 'conocimiento') es un sistema que organiza y construye el conocimiento a través de preguntas comprobables y un método estructurado que estudia e interpreta los fenómeno

-----------------------------------------------------------------

Sample from Colombia.txt:

Colombia, oficialmente República de Colombia, es un país soberano situado en la región noroccidental de América del Sur, actualmente es miembro de la Comunidad Andina y se constituye en un estado

-----------------------------------------------------------------

Sample from Consumo.txt:

Consumo es la acción y efecto de consumir o gastar, sean productos, bienes o servicios, como por ejemplo la energía, entendiendo por consumir, como el hecho de utilizar estos productos y servicios pa

-----------------------------------------------------------------

Sample from Cristianismo.txt:

El cristianismo (del latín christianismus, y este del griego χριστιανισμός)[2]​ es una religión abrahámica monoteísta basada en la vida y enseñanzas de Jesús de Nazaret. Es la religión más extensa de

-----------------------------------------------------------------

Sample from Cristo.txt:

Cristo (del latín Christus, y este del griego antiguo Χριστός, Christós)[1]​ es una traducción del término hebreo «Mesías» (מָשִׁיחַ, Māšîaḥ), que significa «ungido»,[2]​ y que se emplea como título o

-----------------------------------------------------------------

Sample from Cultura.txt:

El término cultura (del latín cultūra)[1]​[2]​ tiene muchos significados interrelacionados, es decir, es un término polisémico. Por ejemplo, en 1952, Alfred Kroeber y Clyde Kluckhohn recopilaron una l

-----------------------------------------------------------------

Sample from Democracia.txt:

La democracia (del latín tardío democratĭa, y este del griego δημοκρατία dēmokratía)[1]​ es una manera de organización social que atribuye la titularidad del poder al conjunto de la ciudadanía. En sen

-----------------------------------------------------------------

Sample from Diccionario.txt:

El diccionario es una obra donde se puede consultar palabras o términos y se proporciona su significado, definición, etimología, ortografía, fija su pronunciación, separación silábica y forma gramatic

-----------------------------------------------------------------

Sample from Divina\_comedia.txt:

La Divina comedia (en italiano moderno: Divina Commedia, en toscano: Divina Comedìa), también conocida simplemente como Comedia, es un poema escrito por Dante Alighieri. Se desconoce la fecha exacta e

-----------------------------------------------------------------

Sample from Ecologismo.txt:

El ecologismo (en ocasiones llamado el movimiento verde o ambientalista) es un variado movimiento político, social y global, que defiende la protección del medio ambiente.[1]​

Habitualmente, el ecolo

-----------------------------------------------------------------

Sample from Energ\_C3\_ADa\_nuclear.txt:

La energía nuclear o atómica es la que se libera espontánea o artificialmente en las reacciones nucleares. Sin embargo, este término engloba otro significado que es el aprovechamiento de dicha energía

-----------------------------------------------------------------

Sample from Enrique\_VIII\_de\_Inglaterra.txt:

Enrique VIII (28 de junio de 1491-28 de enero de 1547) fue rey de Inglaterra y señor de Irlanda desde el 22 de abril de 1509 hasta su muerte. Fue el segundo monarca de la casa Tudor, heredero de su pa

-----------------------------------------------------------------

Sample from Escritura.txt:

La escritura es un sistema de representación gráfica de un idioma, por medio de signos trazados o grabados sobre un soporte. En tal sentido, la escritura es un modo gráfico específicamente humano de

-----------------------------------------------------------------

Sample from Est\_C3\_A9tica.txt:

La estética (del griego αἰσθητική [aesthetic], ‘sensación’, ‘percepción’, y este de[aísthesis], ‘sensación’, ‘sensibilidad’, e -ικά [-icá], ‘relativo a’) es la rama de la filosofía que estudia la esen

-----------------------------------------------------------------

Sample from Estado.txt:

Un Estado es una organización política constituida por un conjunto de instituciones burocráticas estables, a través de las cuales ejerce el monopolio del uso de la fuerza (soberanía) aplicada a una po

-----------------------------------------------------------------

Sample from Estados\_Unidos.txt:

Estados Unidos[nota 2]​ (EE. UU.; en inglés, United States o US),[nota 3]​ cuyo nombre oficial es Estados Unidos de América (EUA; en inglés, United States of America o USA),[12]​ es un país sobera

-----------------------------------------------------------------

Sample from Europa.txt:

Países transcontinentales:

Europa es un continente ubicado enteramente en el hemisferio norte y mayoritariamente en el hemisferio oriental. Las fronteras de Europa están situadas en la mitad occident

-----------------------------------------------------------------

Sample from Filosof\_C3\_ADa.txt:

La filosofía (del griego antiguo φιλοσοφία 'amor a la sabiduría' derivado de φιλεῖν [fileîn] 'amar' y σοφία [sofía] 'sabiduría';[1]​ trans. en latín como philosophĭa)[2]​ es una disciplina académica y

-----------------------------------------------------------------

Sample from Finanzas.txt:

Las finanzas son una rama de la administración y de la economía que estudia el intercambio de capital entre individuos, empresas, o Estados y con la incertidumbre y el riesgo que estas actividades con

-----------------------------------------------------------------

Sample from Fruta.txt:

Se denomina fruta a aquellos frutos comestibles obtenidos de plantas cultivadas o silvestres que, por su sabor generalmente dulce-acidulado, su aroma intenso y agradable y sus propiedades nutritivas,

-----------------------------------------------------------------

Sample from Gabriel\_Garc\_C3\_ADa\_M\_C3\_A1rquez.txt:

Gabriel José García Márquez (Aracataca, 6 de marzo de 1927-Ciudad de México, 17 de abril de 2014)[nota 1]​[2]​ ( escuchar) fue un escritor y periodista colombiano. Reconocido principalmente por sus

-----------------------------------------------------------------

Sample from Gastronom\_C3\_ADa\_de\_M\_C3\_A9xico.txt:

La gastronomía mexicana es el conjunto de platillos y técnicas culinarias de México que forman parte de las tradiciones y vida común de sus habitantes, enriquecida por las aportaciones de las distint

-----------------------------------------------------------------

Sample from George\_Patton.txt:

Tercer Ejército de Estados UnidosDecimoquinto Ejército de Estados Unidos

Expedición Punitiva

Primera Guerra Mundial

Segunda Guerra Mundial

George Smith Patton, Jr. (San Gabriel, California; 11 de

-----------------------------------------------------------------

Sample from Grecia.txt:

Grecia (en griego moderno, Ελλάδα, Elláda, AFI: [eˈlaða] ( escuchar); en griego antiguo, Ἑλλάς, Hellás), oficialmente República Helénica (en griego, Ελληνική Δημοκρατία, Ellinikí Dimokratía, AFI:

-----------------------------------------------------------------

Sample from Guerra\_Fr\_C3\_ADa.txt:

La Guerra Fría fue un enfrentamiento político, económico, social, ideológico, militar e informativo iniciado tras finalizar la Segunda Guerra Mundial entre el bloque Occidental (occidental-capitalista

-----------------------------------------------------------------

Sample from Henry\_Ford.txt:

Henry Ford (Dearborn, Míchigan; 30 de julio de 1863-Ib., 7 de abril de 1947) fue un empresario y emprendedor estadounidense, fundador de la compañía Ford Motor Company y padre de las cadenas de produc

-----------------------------------------------------------------

Sample from Holocausto.txt:

El Holocausto[1]​ —también conocido en hebreo como השואה, Shoá, traducido como «La Catástrofe»—, conocido en la terminología nazi como «solución final» —en alemán, Endlösung— de la «cuestión judía»,[2

-----------------------------------------------------------------

Sample from Homo\_sapiens.txt:

El Homo sapiens (del latín homo, «hombre», y sapiens, «sabio»), hombre o ser humano es una especie del orden de los primates perteneciente a la familia de los homínidos. También son conocidos bajo la

-----------------------------------------------------------------

Sample from Inteligencia\_artificial.txt:

La inteligencia artificial (IA) es, en informática, la inteligencia expresada por máquinas, sus procesadores y sus softwares, que serían los análogos al cuerpo, el cerebro y la mente, respectivamente,

-----------------------------------------------------------------

Sample from Italia.txt:

Italia (en italiano, Italia, pronunciado /i'talja/ ( escuchar)), oficialmente la República Italiana (en italiano, Repubblica Italiana pronunciado /re'pub:lika ita'ljana/), es un país soberano tran

-----------------------------------------------------------------

Sample from Jorge\_Isaacs.txt:

Partido Conservador

Jorge Ricardo Isaacs Ferrer (Santiago de Cali, 1 de abril de 1837-Ibagué, 17 de abril de 1895) fue un novelista, escritor y poeta colombiano del género romántico. Jorge Isaacs vi

-----------------------------------------------------------------

Sample from Juda\_C3\_ADsmo.txt:

El término judaísmo se refiere a la religión, tradición y cultura del pueblo judío. Históricamente, es la más antigua de las tres religiones abrahámicas,[5]​ grupo que tiene como base e incluye el c

-----------------------------------------------------------------

Sample from Juglar.txt:

Un juglar[a]​ —antiguamente también joglar[1]​— era un artista ambulante en la Edad Media. A cambio de dinero o comida, ofrecía su espectáculo callejero en las plazas públicas, y en ocasiones era cont

-----------------------------------------------------------------

Sample from Julio\_Garavito.txt:

Julio Garavito Armero (Bogotá, 5 de enero de 1865-Bogotá, 11 de marzo de 1920) fue un astrónomo, matemático, economista, poeta e ingeniero colombiano.[1]​ Sus investigaciones contribuyeron al desarrol

-----------------------------------------------------------------

Sample from Liberalismo.txt:

El liberalismo es una filosofía y movimiento político que defiende la libertad individual, la igualdad ante la ley y la limitación de los poderes del Estado. Como doctrina económica propugna la inicia

-----------------------------------------------------------------

Sample from Libro.txt:

Un libro (del latín liber, libri) es una obra impresa, manuscrita o pintada en una serie de hojas de papel, pergamino, vitela u otro material, unidas por un lado (es decir, encuadernadas) y protegidas

-----------------------------------------------------------------

Sample from Ludwig\_van\_Beethoven.txt:

Ludwig van Beethoven[a]​ (Bonn, Arzobispado de Colonia; 16 de diciembre de 1770[b]​-Viena, 26 de marzo de 1827) fue un compositor, director de orquesta, pianista y profesor de piano alemán. Su legado

-----------------------------------------------------------------

Sample from Nevado\_del\_Ruiz.txt:

El Volcán del Ruiz , Nevado del ruiz , también conocido como Mesa de Herveo,[4]​ y en la época precolombina como Cumanday, Tabuchía y Tama,[1]​ se conoce como Volcán del Ruiz quizás en honor a Alfonso

-----------------------------------------------------------------

Sample from Nikola\_Tesla.txt:

Nikola Tesla (en serbio, Никола Тесла; Smiljan, Imperio austríaco, actual Croacia; 10 de julio de 1856-Nueva York, 7 de enero de 1943) fue un inventor, ingeniero eléctrico y mecánico serbio nacional

-----------------------------------------------------------------

Sample from Odisea.txt:

La Odisea (en griego: Ὀδύσσεια, Odýsseia) es un poema épico griego compuesto por 24 cantos, atribuido al poeta griego Homero. Se cree que fue compuesta en el siglo VIII a. C. en los asentamientos que

-----------------------------------------------------------------

Sample from Organizaci\_C3\_B3n\_de\_las\_Naciones\_Unidas.txt:

La Organización de las Naciones Unidas (ONU), también conocida simplemente como Naciones Unidas (NN. UU.), es la mayor organización internacional existente. Se creó para mantener la paz y seguridad in

-----------------------------------------------------------------

Sample from Per\_C3\_ADodo\_helen\_C3\_ADstico.txt:

Se denomina período helenístico, helenismo o periodo alejandrino (por Alejandro Magno) a una etapa histórica de la Antigüedad cuyos límites cronológicos vienen marcados por dos importantes acontecimie

-----------------------------------------------------------------

Sample from Pizza.txt:

La pizza es una preparación culinaria que consiste en un pan plano, habitualmente de forma circular, elaborado con harina de trigo, levadura, agua y sal (a veces aceite de oliva) que tradicionalmente

-----------------------------------------------------------------

Sample from Pueblos\_germ\_C3\_A1nicos.txt:

Los pueblos germánicos o germanos son un histórico grupo etnolingüístico de los pueblos originarios del norte de Europa que se identifican por el uso de las lenguas germánicas (un subgrupo de la fam

-----------------------------------------------------------------

Sample from Real\_Academia\_Espa\_C3\_B1ola.txt:

La Real Academia Española (RAE)[3]​ es una institución cultural española privada financiada con fondos públicos con sede en Madrid, España. Esta y otras veintitrés academias de la Lengua correspondien

-----------------------------------------------------------------

Sample from Reino\_Unido.txt:

El Reino Unido (United Kingdom en inglés),[nota 1]​ o de forma abreviada RU (UK en inglés), oficialmente Reino Unido de Gran Bretaña e Irlanda del Norte (United Kingdom of Great Britain and Northe

-----------------------------------------------------------------

Sample from Revoluci\_C3\_B3n\_Industrial.txt:

La Revolución Industrial o Primera Revolución Industrial es el proceso de transformación económica, social y tecnológica que se inició en la segunda mitad del siglo XVIII en el Reino de Gran Bretaña,

-----------------------------------------------------------------

Sample from Rueda.txt:

La rueda es un elemento circular y mecánico que gira alrededor de un eje.[1]​ Puede ser considerada una máquina simple, y forma parte del conjunto denominado elementos de máquinas.

Es uno de los inve

-----------------------------------------------------------------

Sample from Salsa\_(g\_C3\_A9nero\_musical).txt:

La salsa es un género musical bailable resultante de la síntesis del son cubano y otros géneros de música caribeña y estadounidenses como el jazz y el blues. La salsa se consolidó como un éxito comer

-----------------------------------------------------------------

Sample from Saxof\_C3\_B3n.txt:

El saxofón, también conocido como saxófono o simplemente saxo, es un instrumento musical cónico, de la familia de los instrumentos de viento-madera, generalmente hecho de latón, que consta de una boqu

-----------------------------------------------------------------

Sample from Segunda\_Guerra\_Mundial.txt:

Invasión alemana de Polonia

 Alemania nazi

Reino de Italia

 Imperio del Japón

Francia de Vichy

Reino de Hungría

Reino de Rumania

Albania

Eslovaquia

Reino de Bulgaria

Croacia

Mengjiang

Manc

-----------------------------------------------------------------

Sample from Sublime.txt:

Lo sublime es una categoría estética, derivada principalmente de la célebre obra Περὶ ὕψους (Sobre lo sublime) del crítico o retórico griego Longino (o Pseudo-Longino), y que consiste fundamentalmente

-----------------------------------------------------------------

Sample from Thomas\_Alva\_Edison.txt:

Thomas Alva Edison (Milan, Ohio; 11 de febrero de 1847-West Orange, Nueva Jersey; 18 de octubre de 1931) fue un inventor, científico y empresario estadounidense.[1]​[2]​[3]​ Desarrolló numerosos dispo

-----------------------------------------------------------------

Sample from Valle.txt:

Un valle (del latín vallis) es una llanura entre montañas o alturas, una depresión de la superficie terrestre entre dos vertientes, con forma inclinada y alargada, que conforma una cuenca hidrográfica

-----------------------------------------------------------------

**Verificación final de los datos:**

En el siguiente enlace se podrá encontrar el documento desarrollado para realizar la extracción, procesamiento y análisis de los datos de los distintos archivos. Lo expuesto en dicho documento constituye el detalle de cómo se realizó el tratamiento de los datos y su análisis, a describir brevemente en los siguientes puntos:

<https://github.com/amendivilsejin/training_analitica_2nd_module_repo/blob/master/Data_understanding_exercises.ipynb>

A partir del tratamiento de los datos, se lograron extraer 32430 palabras. Una ilustración de la frecuencia con la que se encontraron las palabras en las distintas fuentes de datos, se presenta en el siguiente mapa de palabras. En este, el tamaño de cada palabra es proporcional a su frecuencia en el conjunto de datos:



Se procesaron y filtraron los archivos con acompañamiento del cliente, de acuerdo a lineamientos provistos por el mismo. Se verificó el resultado de la aplicación de los lineamientos de limpieza con un banco final de referencia. Se encontró un resultado satisfactorio del proceso de limpieza: es decir, todas las palabras encontradas se habían previamente reconocido por el cliente en el banco de datos, de manera idéntica.

**Medidas de tendencia central, posicionamiento, y dispersión estadística:**

El conjunto original de palabras extraído de los datos, fue utilizado para construir un conjunto de palabras únicas. A partir de este conjunto, se pudieron extraer algunas estadísticas, por ejemplo, relativas a las longitudes de las palabras:

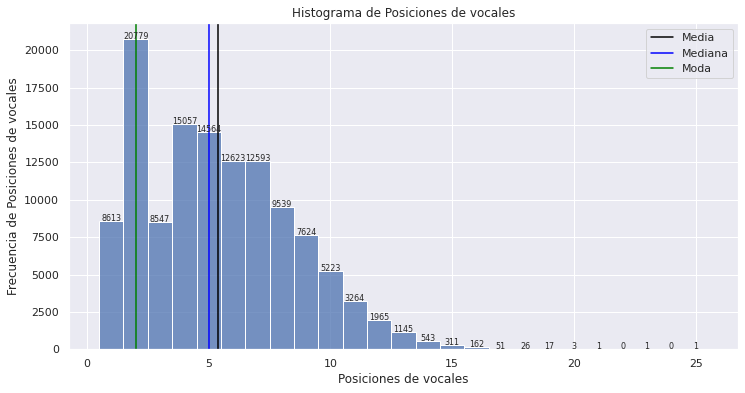
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Parámetro** | **Media de parám.** | **Varianza de parám.** | **Desv. Estándar.** | **Moda de parám.** | **Rango de parám.** | **Cuartil 1 de parám.** | **Mediana de parám** | **Cuartil 3 de parám.** | **Percentil 90  de parám.** |
| **Longitudes de palabras** | **8.50** | **6.78** | **2.60** | **8.0** | **21 [4 a 25] caracteres** | **7.0** | **8.0** | **10.0** | **12.0** |

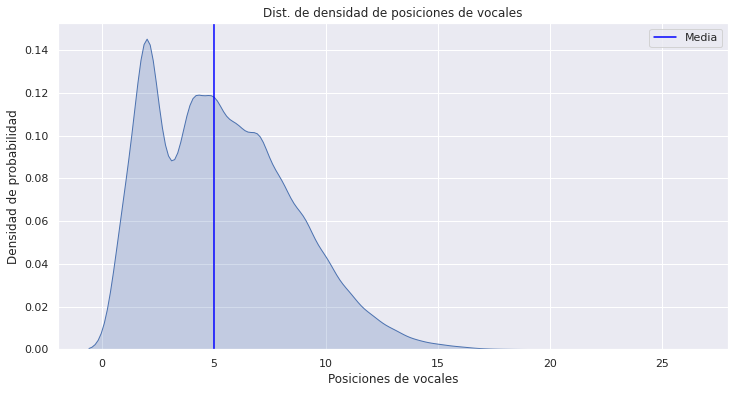
O, por otro lado, relativas a las posiciones de cada una de las letras del alfabeto, en todas las palabras en las que se encontrasen:

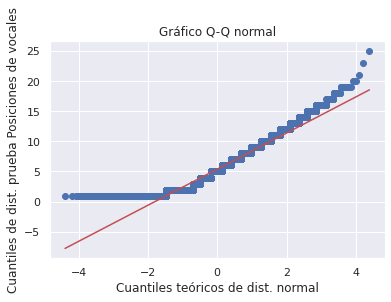
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Letra** | **Frecuencia de letra** | **Promedio de posiciones** | **Varianza de posiciones** | **Desviación estándar de posiciones** | **Moda de posiciones** | **Rango de posiciones** | **Cuartil 1 de dist. De posiciones** | **Mediana de posiciones** | **Cuartil 3 de dist. De posiciones** |
| a | 35862 | 5.62 | 9.53 | 3.09 | 2 | 22 | 3.0 | 6.0 | 8.0 |
| e | 27468 | 4.89 | 10.07 | 3.17 | 2 | 18 | 2.0 | 4.0 | 7.0 |
| i | 26036 | 5.44 | 7.60 | 2.76 | 4 | 20 | 4.0 | 5.0 | 7.0 |
| o | 25322 | 5.99 | 10.63 | 3.26 | 2 | 24 | 3.0 | 6.0 | 8.0 |
| r | 21340 | 4.77 | 6.95 | 2.64 | 3 | 15 | 3.0 | 5.0 | 6.0 |
| n | 20218 | 5.93 | 9.92 | 3.15 | 3 | 23 | 3.0 | 6.0 | 8.0 |
| s | 19517 | 6.31 | 13.51 | 3.68 | 3 | 20 | 3.0 | 6.0 | 9.0 |
| t | 14964 | 5.37 | 8.55 | 2.92 | 4 | 21 | 3.0 | 5.0 | 7.0 |
| c | 14735 | 4.48 | 9.01 | 3.00 | 1 | 16 | 1.0 | 4.0 | 7.0 |
| d | 11743 | 5.52 | 10.57 | 3.25 | 1 | 20 | 3.0 | 6.0 | 8.0 |
| l | 11361 | 4.84 | 7.42 | 2.72 | 3 | 19 | 3.0 | 4.0 | 6.0 |
| m | 8130 | 4.13 | 8.17 | 2.86 | 1 | 19 | 2.0 | 3.0 | 6.0 |
| u | 7964 | 3.98 | 4.54 | 2.13 | 2 | 16 | 2.0 | 4.0 | 5.0 |
| p | 6825 | 2.76 | 4.05 | 2.01 | 1 | 17 | 1.0 | 3.0 | 4.0 |
| b | 4648 | 4.02 | 7.04 | 2.65 | 1 | 16 | 2.0 | 3.0 | 6.0 |
| g | 4414 | 4.07 | 5.65 | 2.38 | 3 | 15 | 3.0 | 4.0 | 6.0 |
| v | 3385 | 4.10 | 7.22 | 2.69 | 1 | 14 | 2.0 | 4.0 | 6.0 |
| f | 3072 | 3.33 | 5.87 | 2.42 | 1 | 16 | 1.0 | 3.0 | 5.0 |
| h | 2437 | 3.26 | 5.95 | 2.44 | 1 | 14 | 1.0 | 3.0 | 5.0 |
| z | 1642 | 6.07 | 6.94 | 2.63 | 7 | 14 | 4.0 | 6.0 | 8.0 |
| j | 1082 | 3.45 | 5.02 | 2.24 | 1 | 13 | 1.0 | 3.0 | 5.0 |
| y | 816 | 4.74 | 5.79 | 2.41 | 3 | 13 | 3.0 | 5.0 | 6.0 |
| k | 763 | 3.93 | 6.83 | 2.61 | 1 | 14 | 1.0 | 4.0 | 5.0 |
| x | 725 | 2.78 | 2.41 | 1.55 | 2 | 11 | 2.0 | 2.0 | 3.0 |
| q | 619 | 3.47 | 4.30 | 2.07 | 3 | 13 | 2.0 | 3.0 | 5.0 |
| w | 369 | 3.00 | 4.84 | 2.20 | 1 | 11 | 1.0 | 3.0 | 5.0 |
| ñ | 280 | 4.88 | 3.86 | 1.97 | 3 | 12 | 3.0 | 5.0 | 6.0 |

**Pruebas de hipótesis:**

También se analizaron variables como la posición de las vocales, y la posición de las consonantes. Inicialmente, se compararon sus distribuciones:





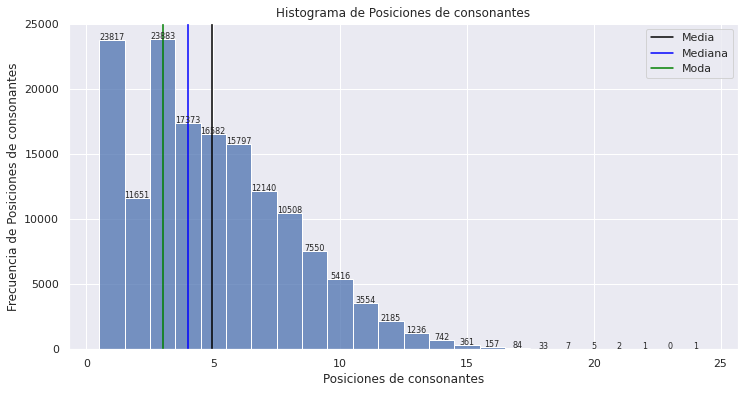


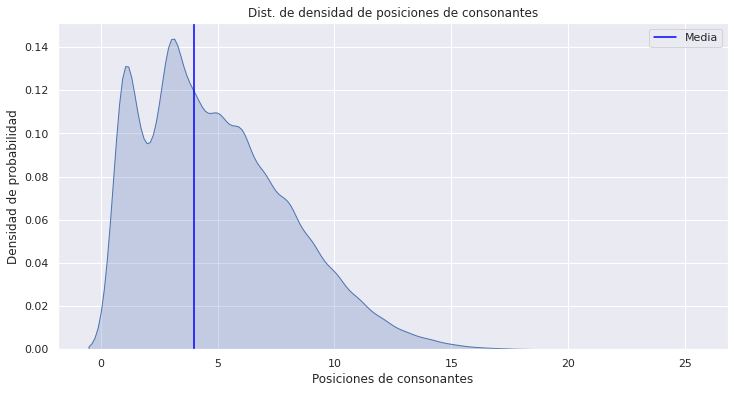
Test de Shapiro-Wilk W: 0.949, p-value: 0.0

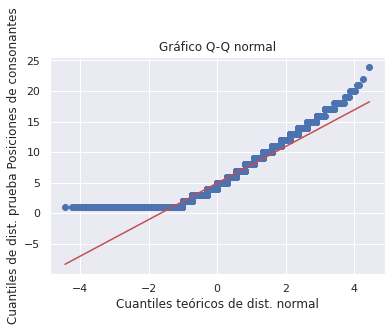
Sesgo: 0.584

Curtosis: -0.097

Concluimos que la distribución de las posiciones de las vocales en las palabras no es normal. Ahora, analicemos la distribución de las posiciones de las consonantes:







Test de Shapiro-Wilk W: 0.937, p-value: 0.0

Sesgo: 0.712

Curtosis: 0.111

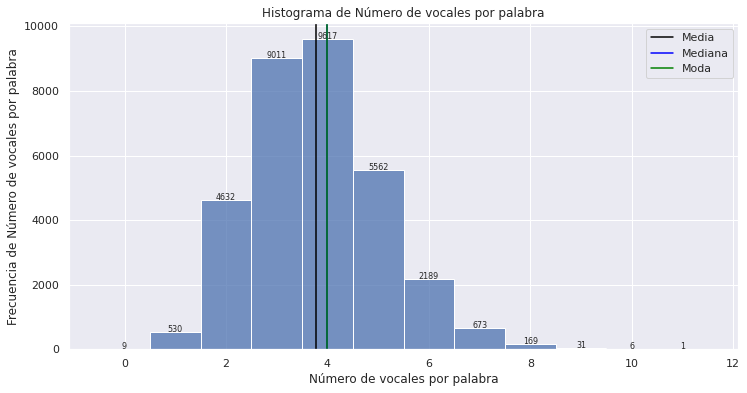
Concluimos que la distribución de las posiciones de las consonantes en las palabras no es normal. A partir de los resultados anteriores, podemos revisar si las distribuciones de las variables analizadas son homogéneas. Para esto, utilizamos una prueba de Chi cuadrado, de 2 muestras independientes:

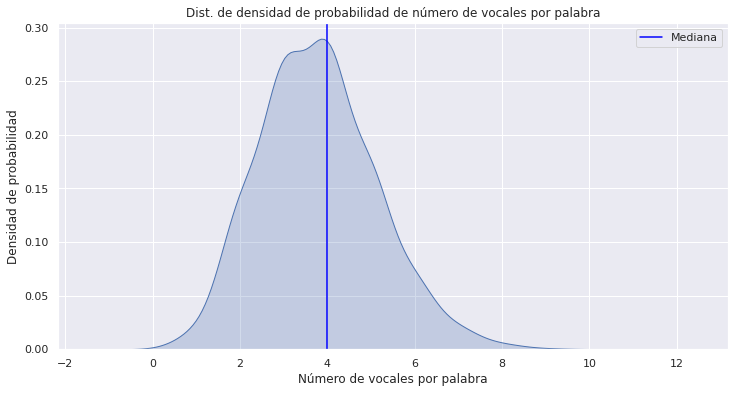
Chi2: 14535.442484437363, p-value: 0.0, DOF: 12

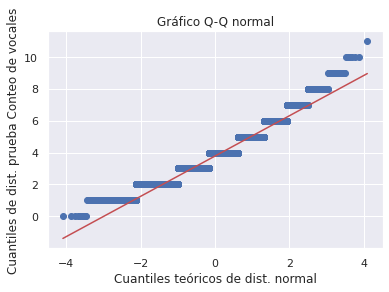
A partir de la prueba de hipótesis realizada anteriormente, concluimos que las distribuciones de las posiciones de consonantes y vocales son distintas.

**Determinación de correlaciones entre variables:**

Por otro lado, se analizó el comportamiento entre las variables Cantidad de vocales por palabra, Cantidad de consonantes por palabra, y Longitud de las palabras:





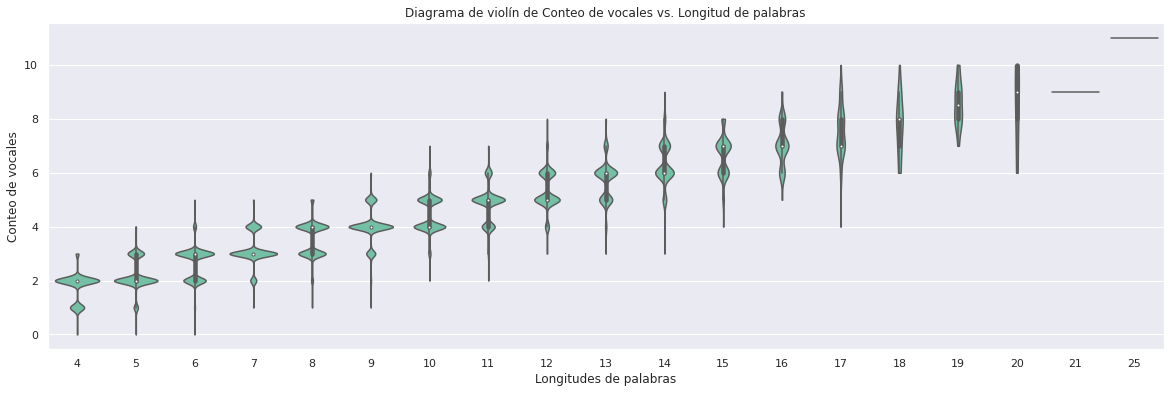


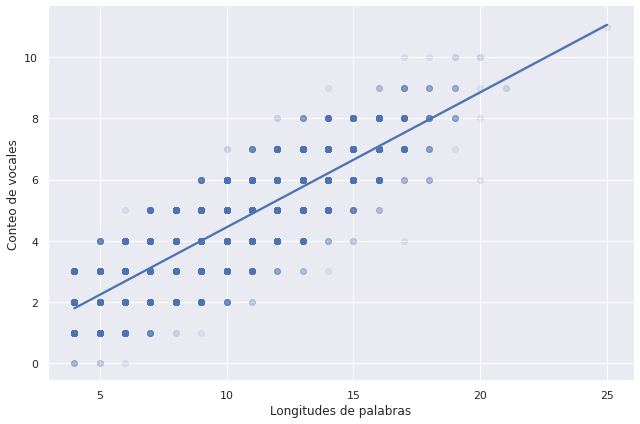
Test de Shapiro-Wilk W: 0.937, p-value: 0.0

Sesgo: 0.459

Curtosis: 0.311

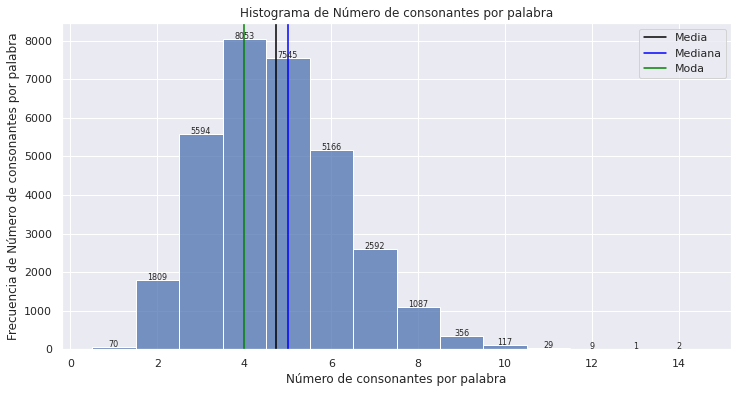
Concluimos que la distribución de las vocales en las palabras no es normal. Ahora, analizamos el comportamiento de esta variable, frente al comportamiento de la variable longitud de palabras:

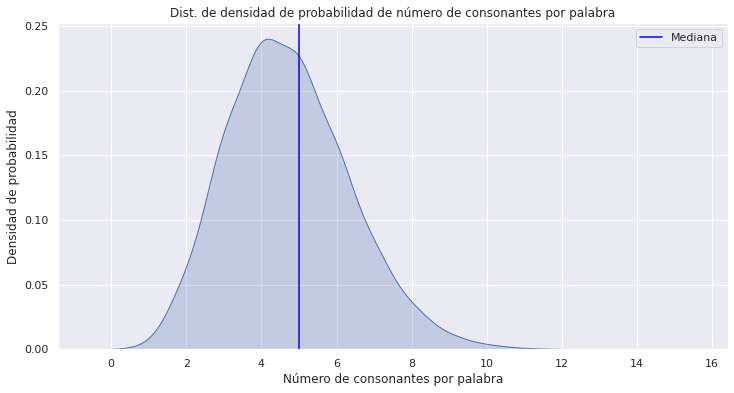


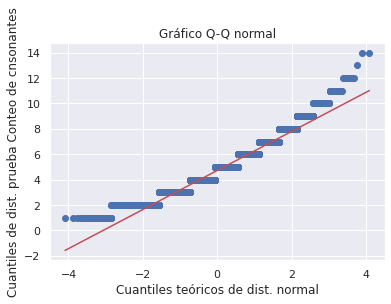


Encontramos que el número de vocales en cada palabra, comparada con su longitud, sigue una tendencia lineal positiva fuerte. Más adelante, podríamos cuantificar esta correlación

Ahora, analizamos la distribución de las consonantes en las palabras:





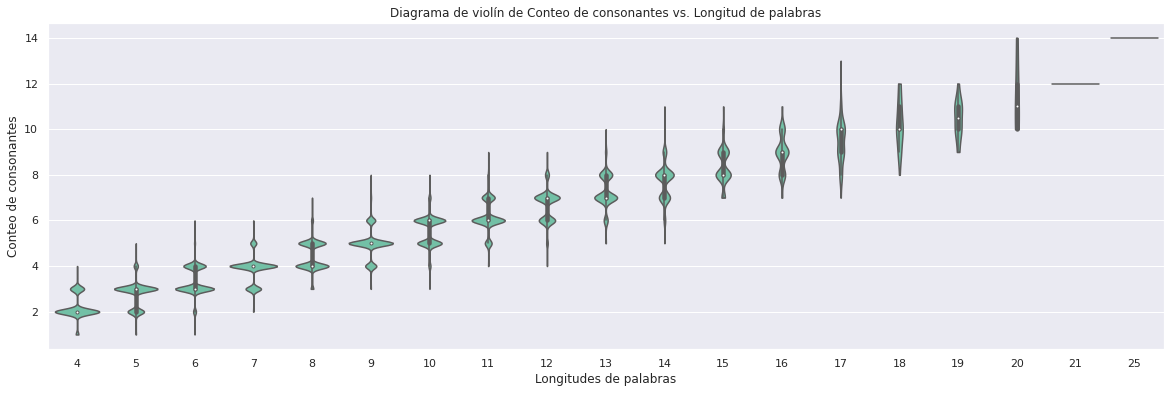


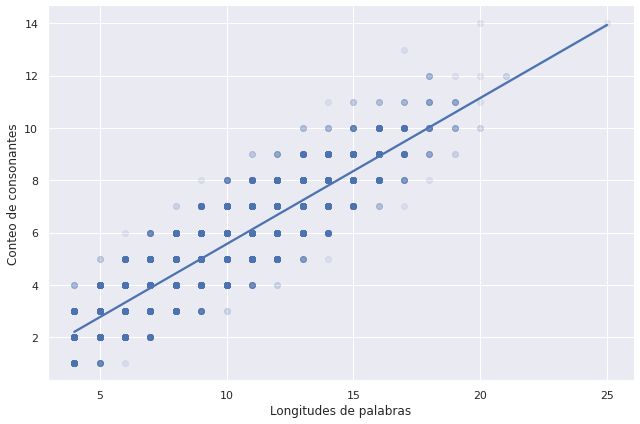
Test de Shapiro-Wilk W: 0.95, p-value: 0.0

Sesgo: 0.52

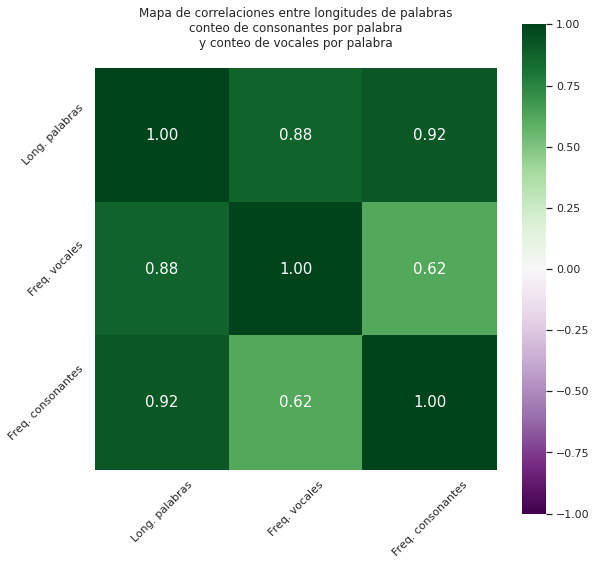
Curtosis: 0.334

Concluimos que la distribución de las consonantes en las palabras no es normal. Ahora, analizamos el comportamiento de esta variable, frente al comportamiento de la variable longitud de palabras:

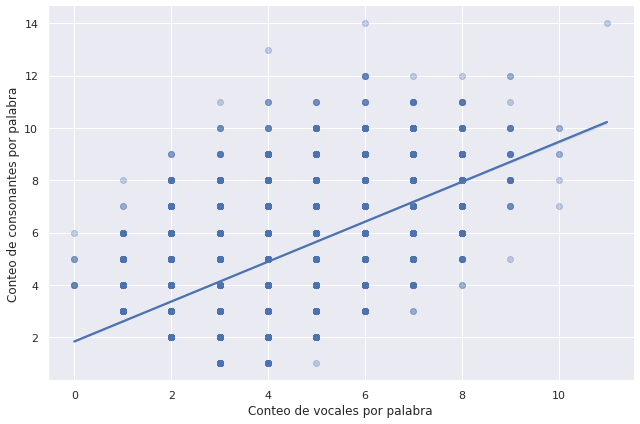




Encontramos que el número de consonantes en cada palabra, comparada con su longitud, sigue una tendencia lineal positiva fuerte. Ahora, cuantifiquemos las correlaciones identificadas entre las variables analizadas:



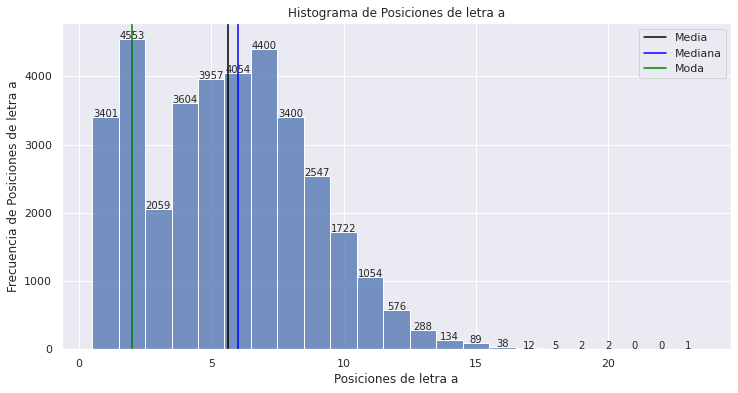
El diagrama anterior presenta las correlaciones de Spearman entre las variables cantidad de vocales, cantidad de consonantes, y longitud de palabras. Se evidencia que la cantidad de vocales y la longitud de palabras, así como la cantidad de consonantes y la longitud de palabras, siguen una correlación positiva y fuerte. Otro hallazgo es que la cantidad de vocales, y la cantidad de consonantes por palabra también presentan correlación positiva, pero esta es más débil. La relación entre estas variables puede entenderse mejor a partir de la siguiente gráfica:

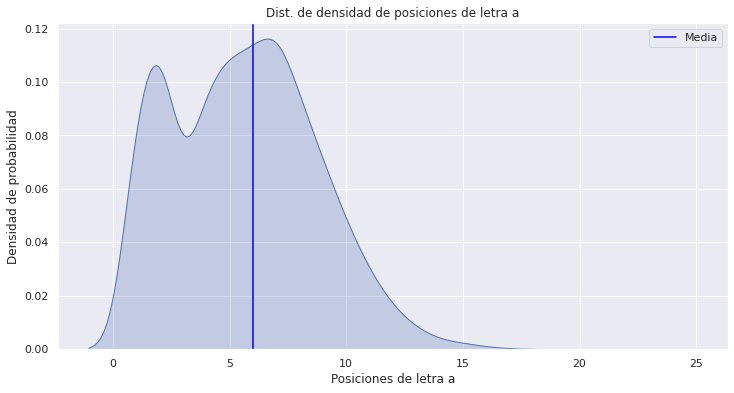


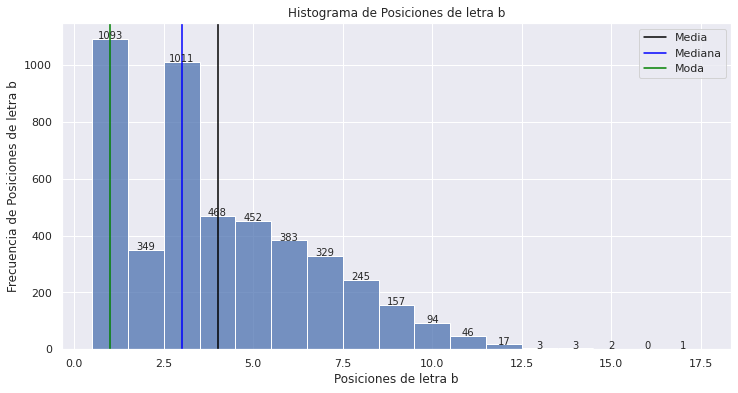
Esto lo que significa es que el radio de consonantes/vocales, aunque creciente, de acuerdo a la longitud de las palabras, presenta mucha variabilidad en el rango de palabras con longitud intermedia (cercana a la media).

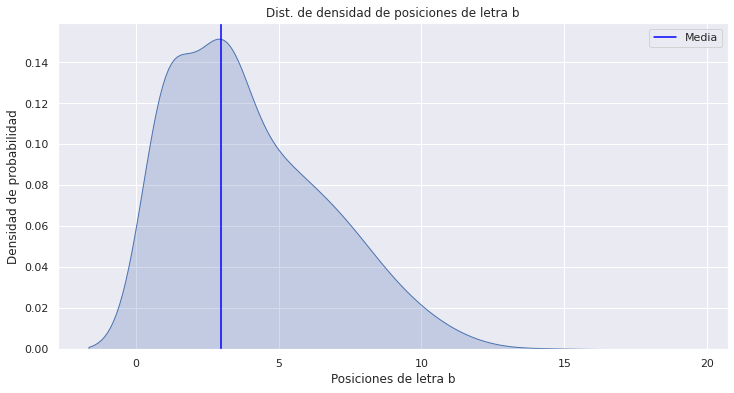
**Distribuciones de datos:**

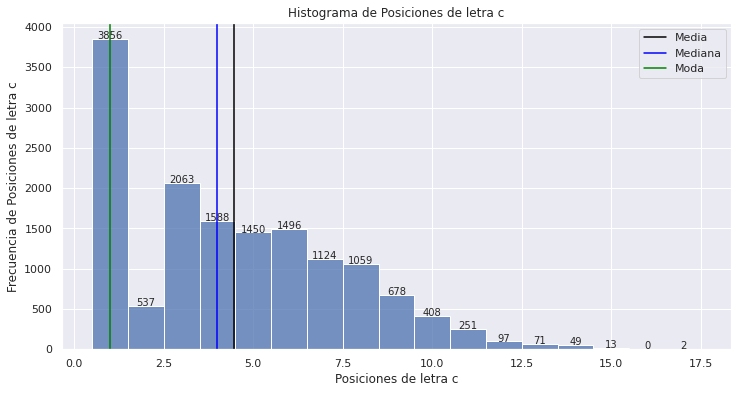
También se revisaron las distribuciones de variables como la longitud de las palabras, o variables como la posición de cada letra del alfabeto en todas las palabras del banco de palabras únicas hallado:

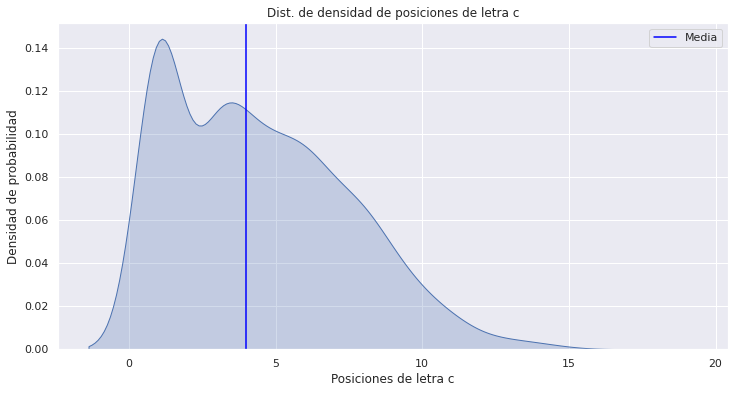


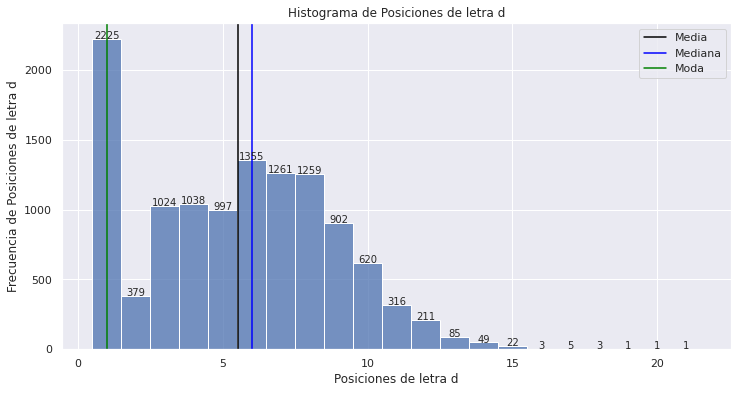


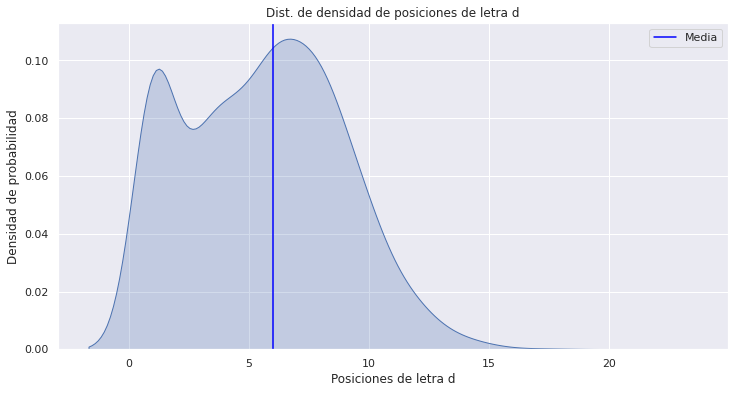


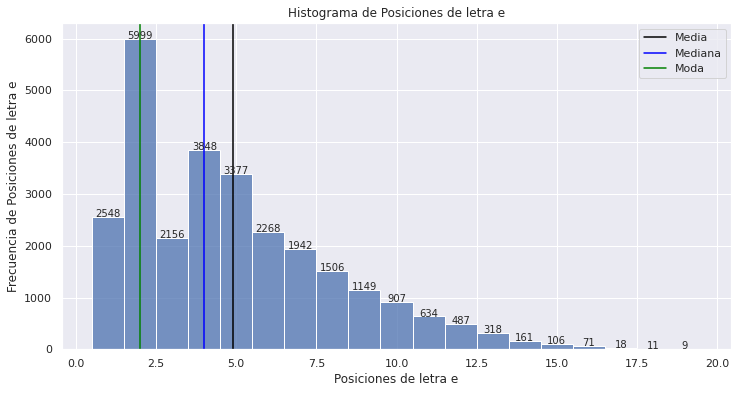


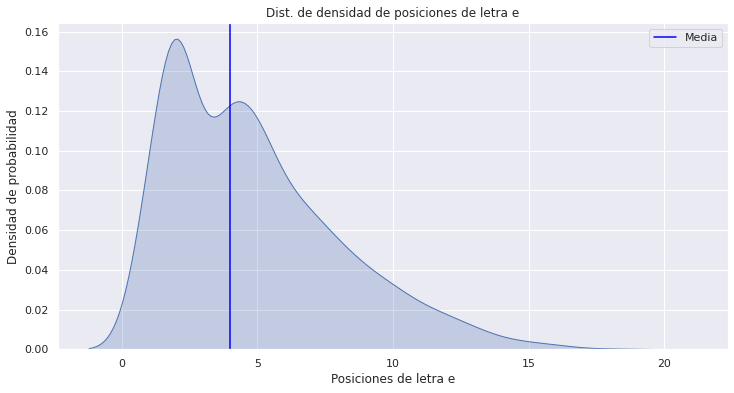


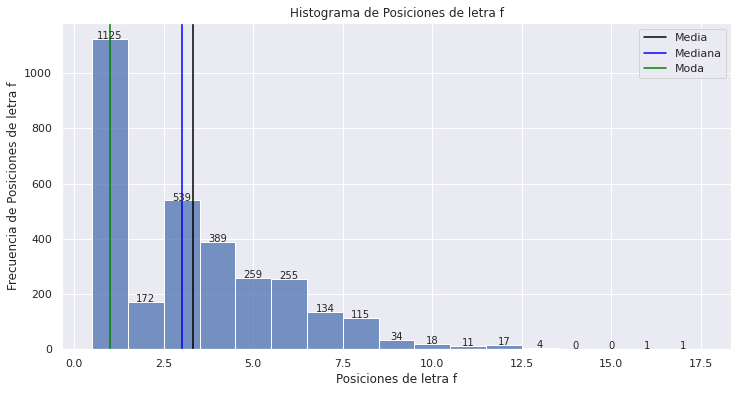


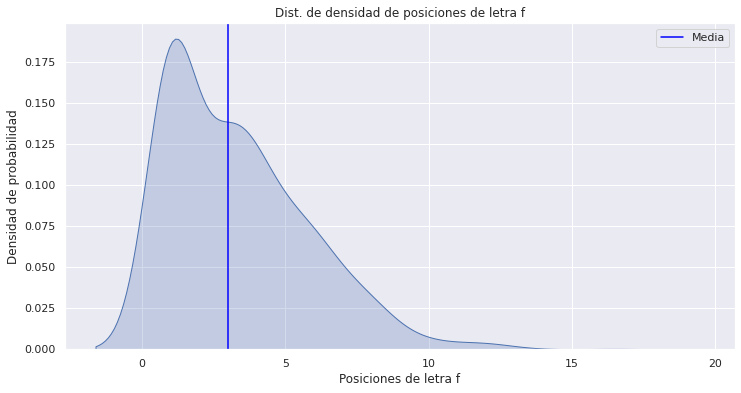


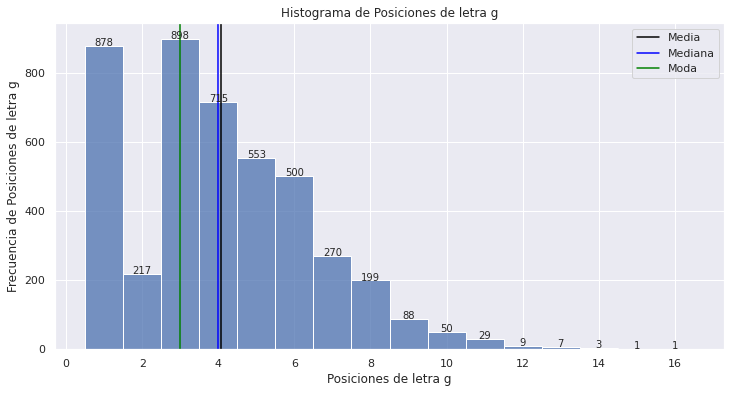


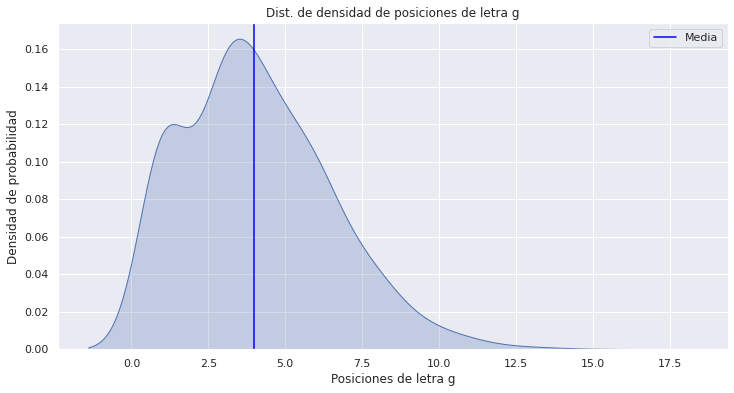


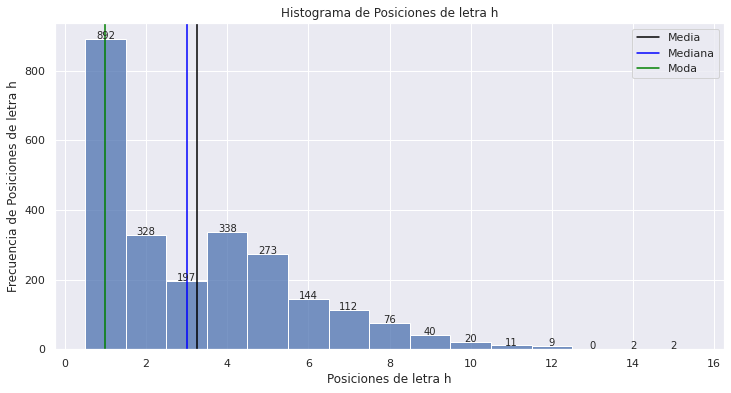


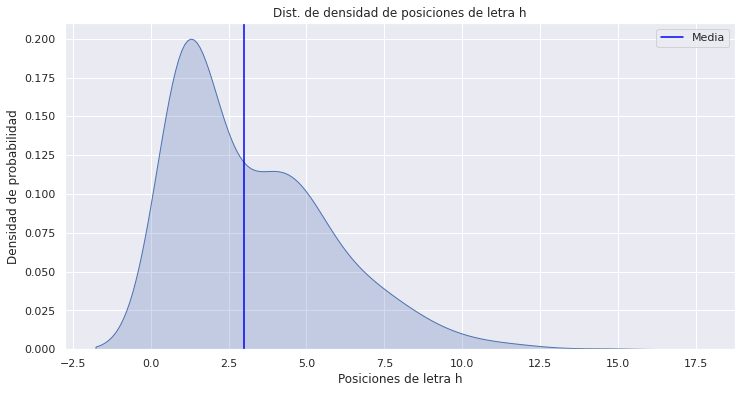


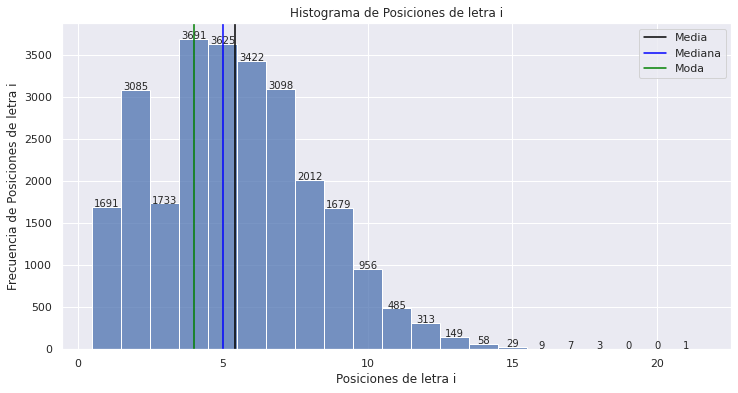


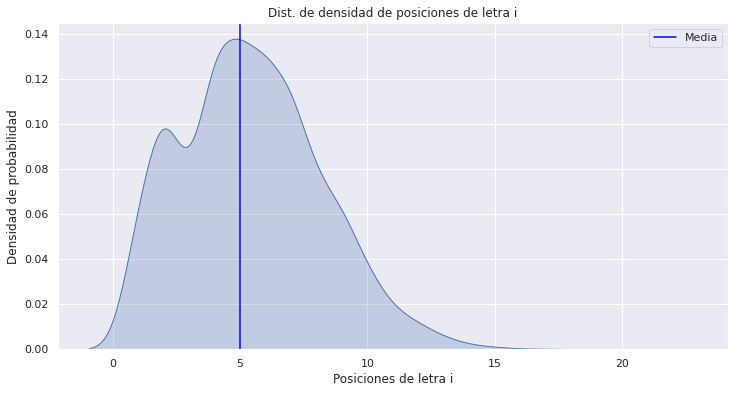


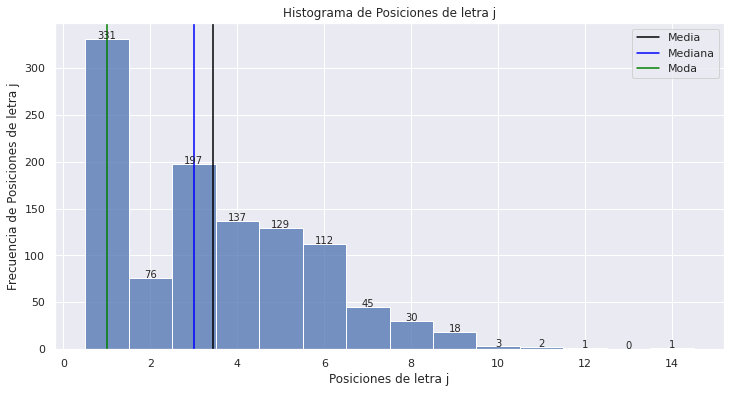


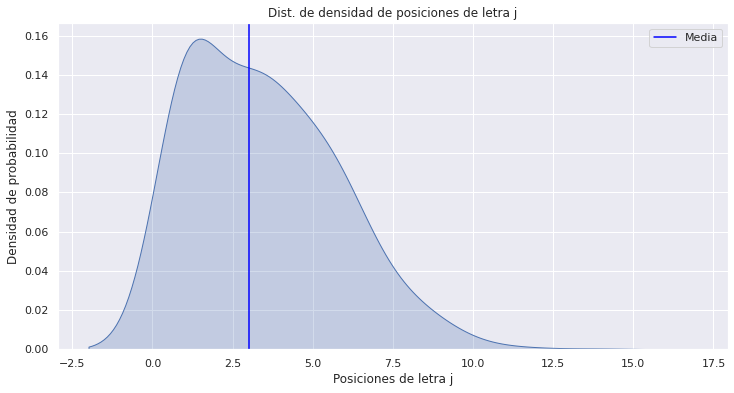


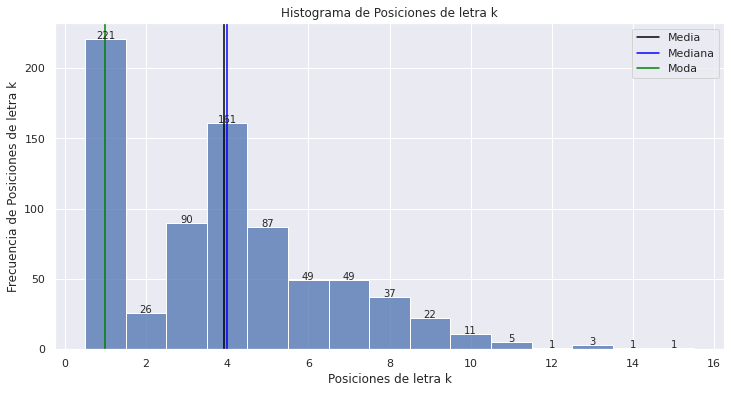


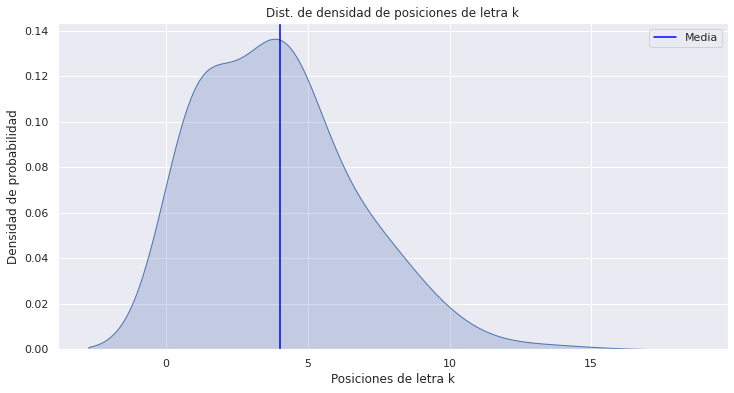


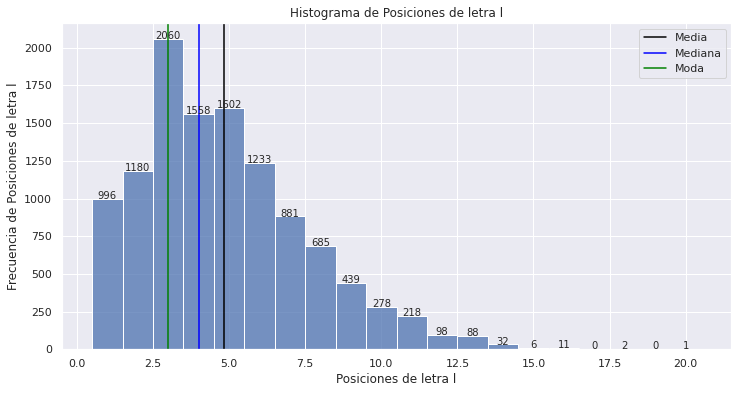


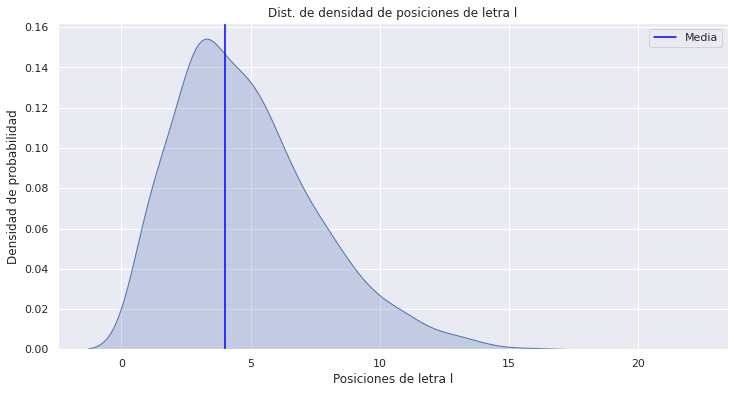


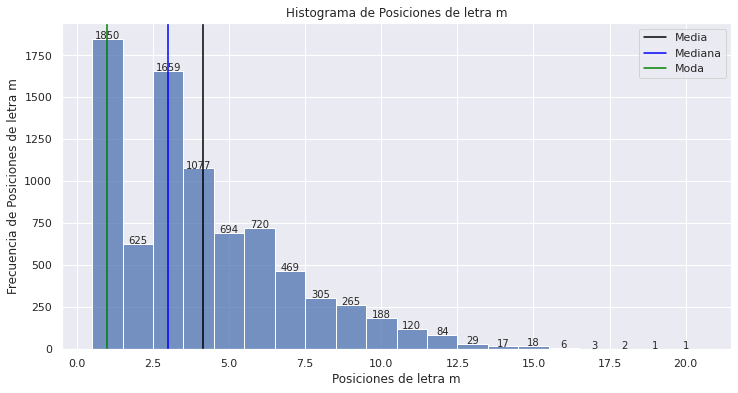


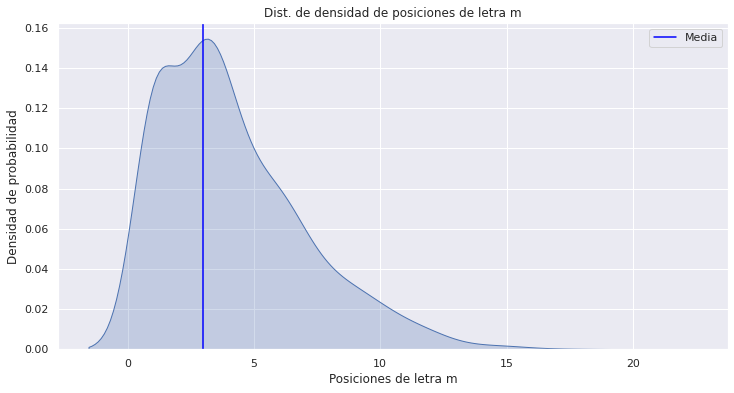


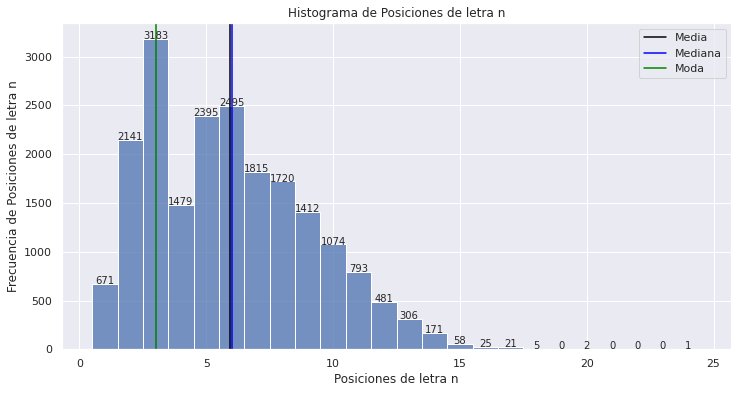


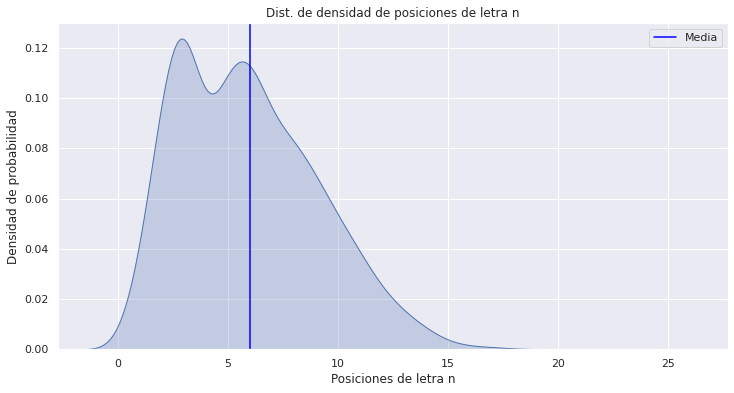


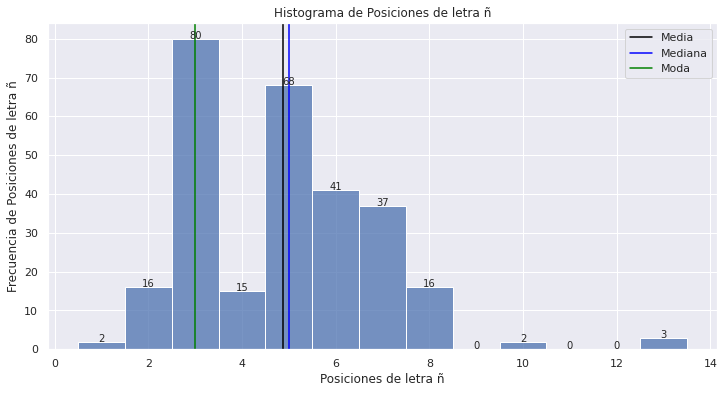


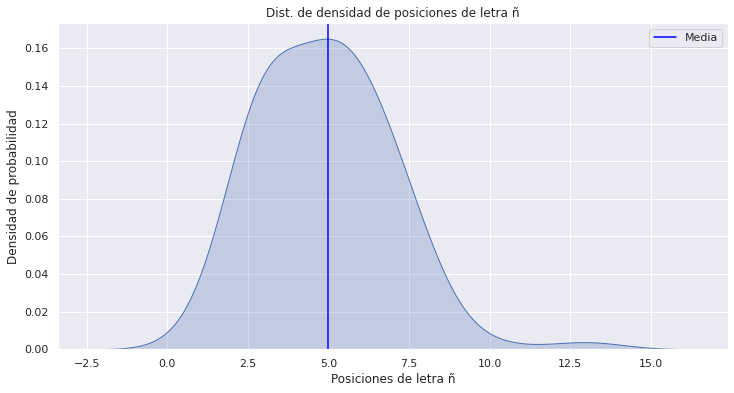


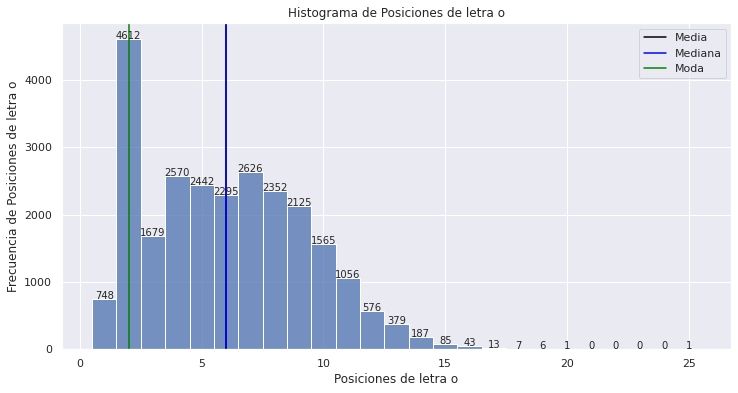


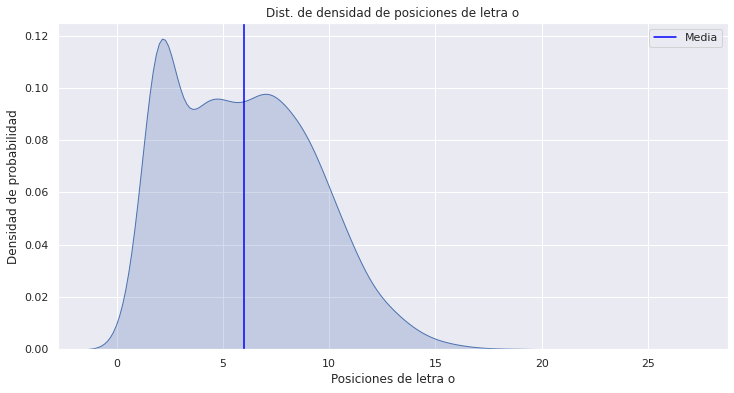


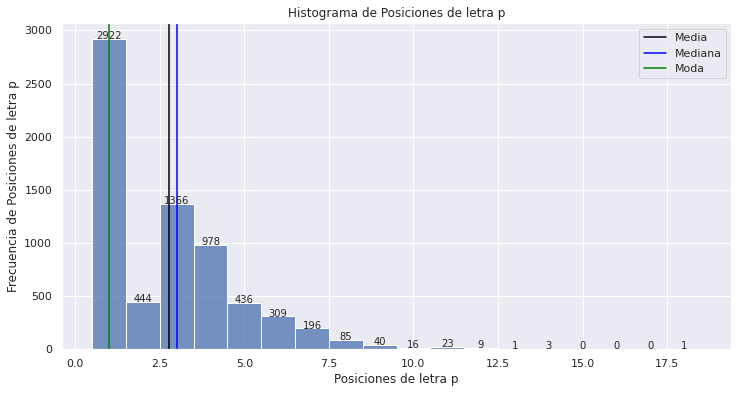


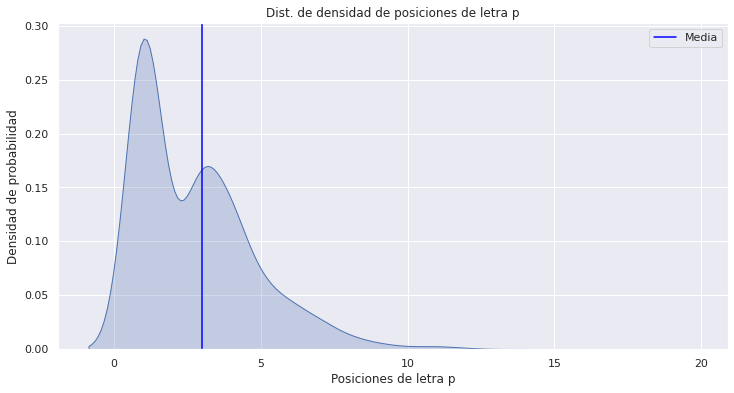


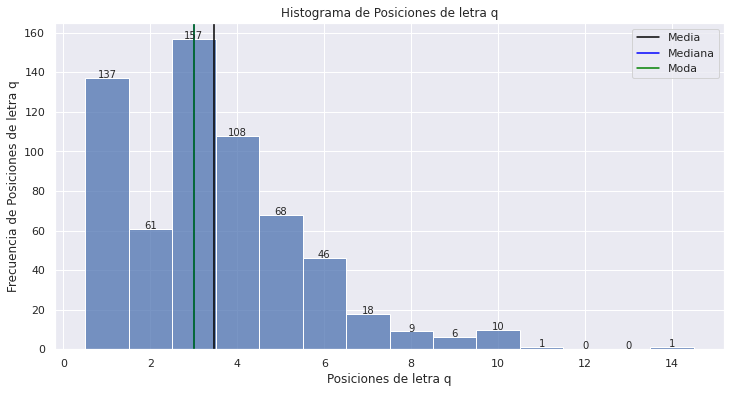


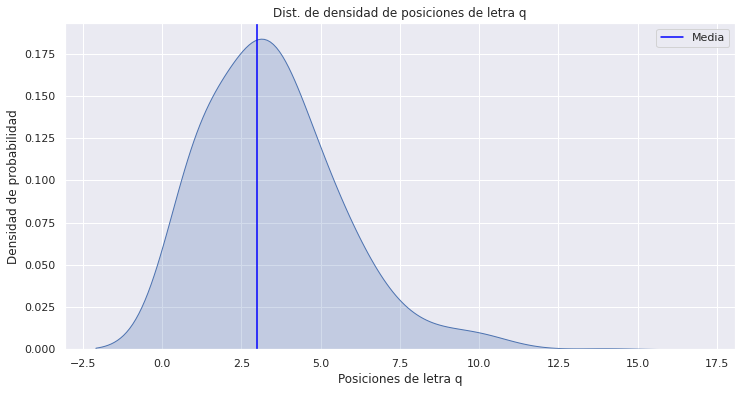


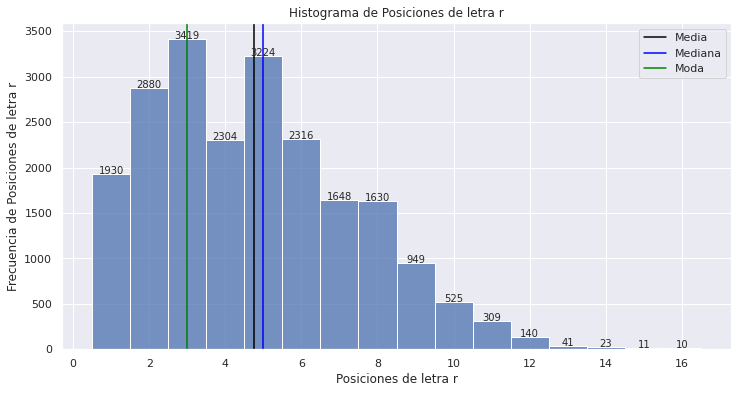


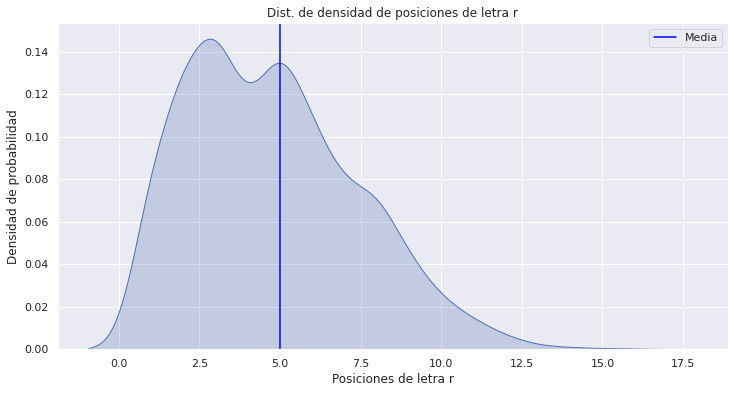


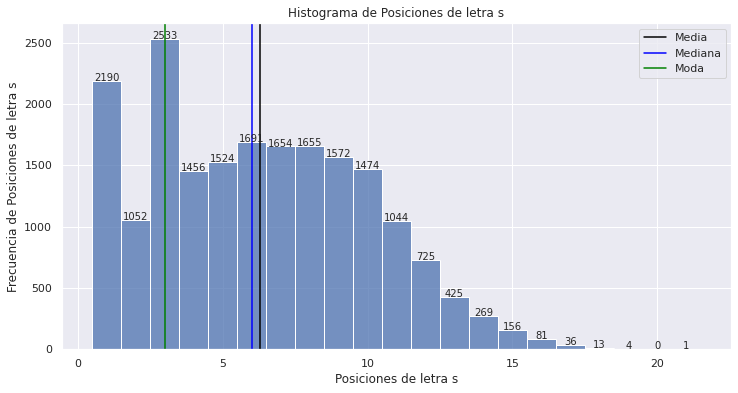


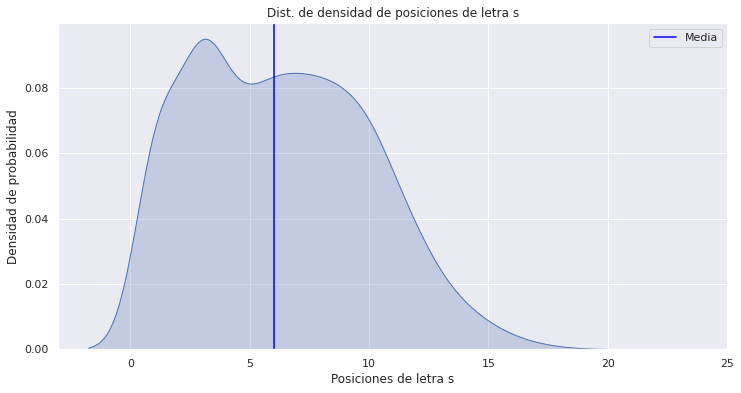


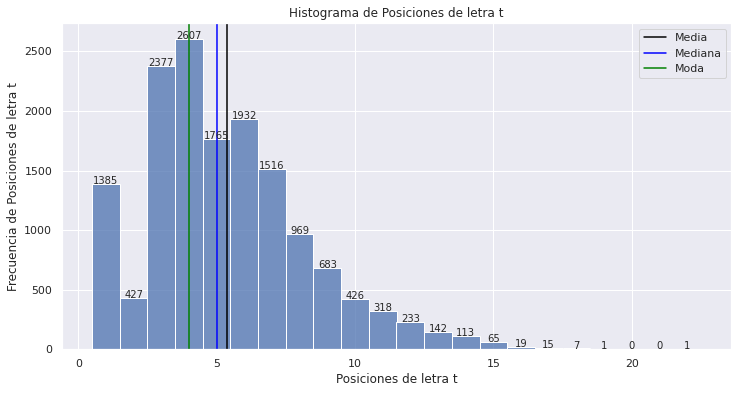


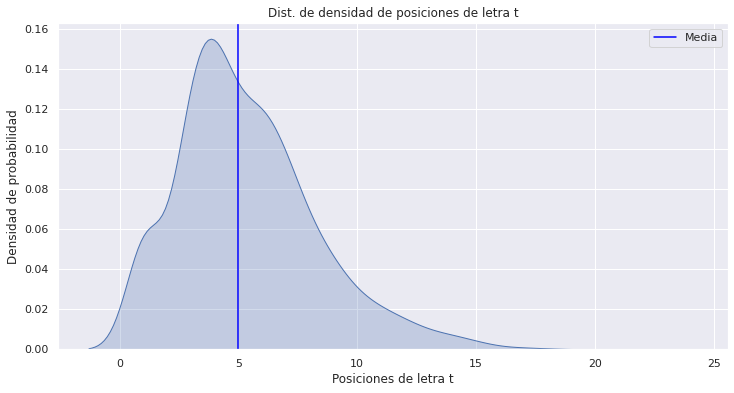


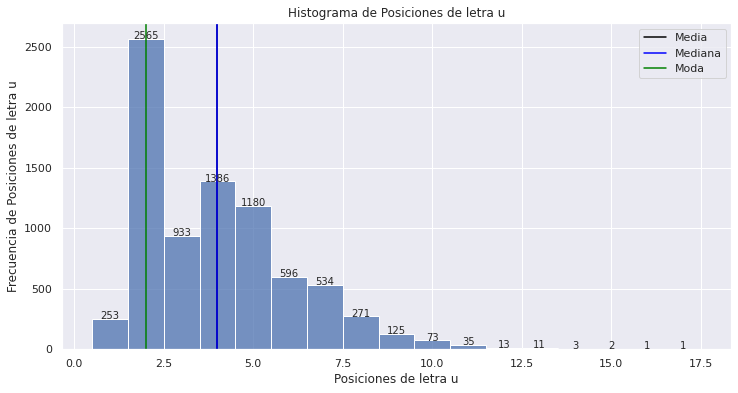


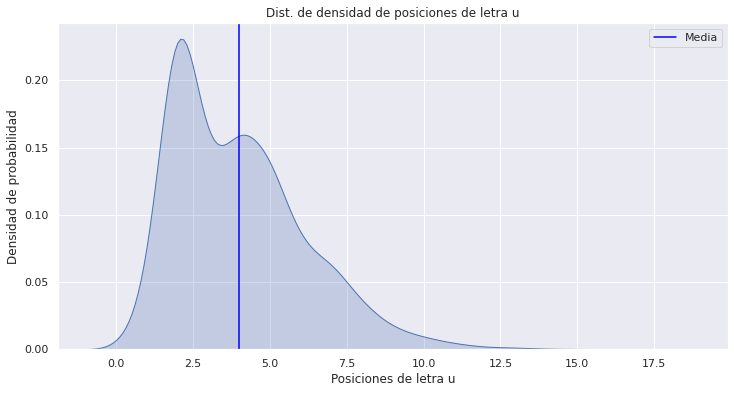


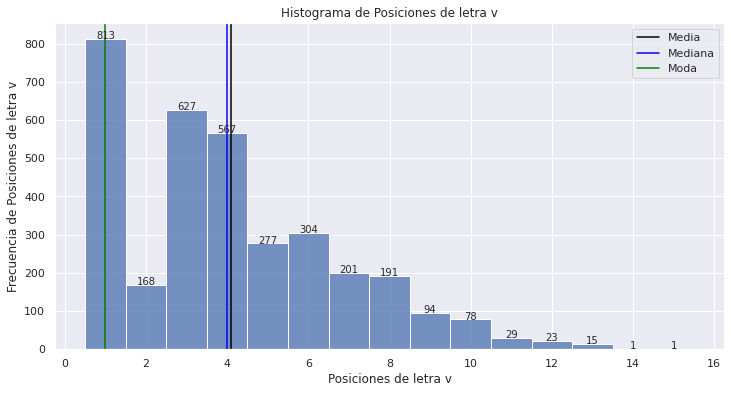


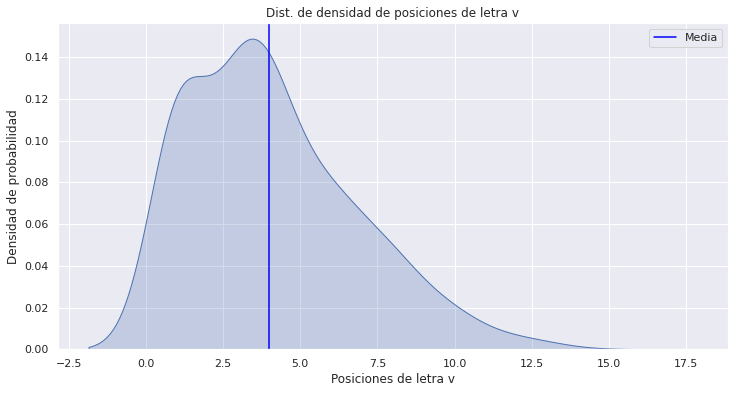


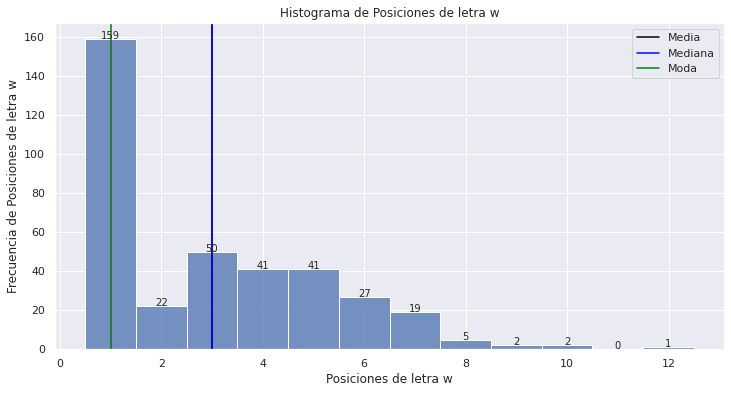


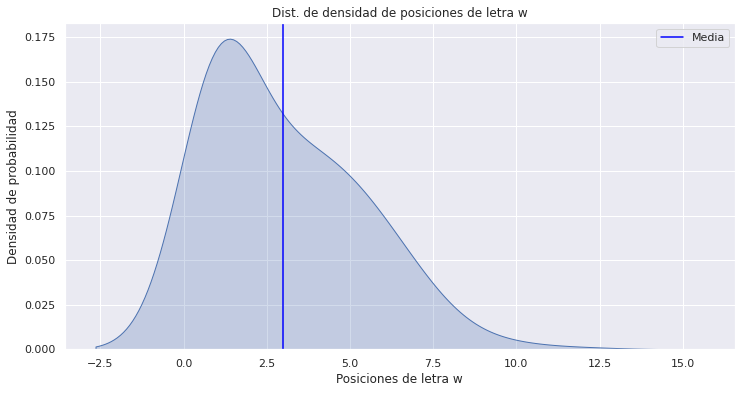


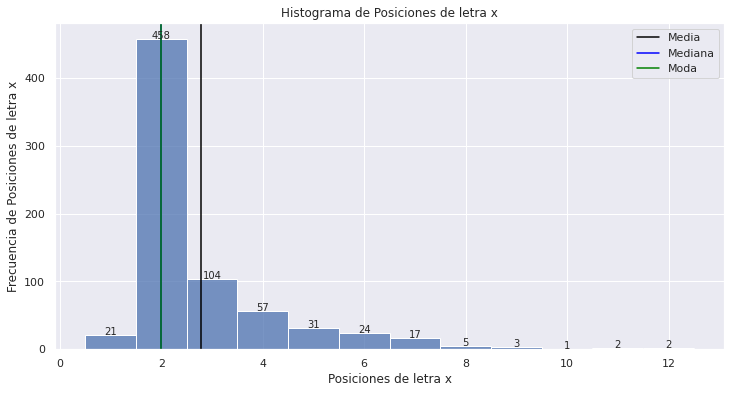


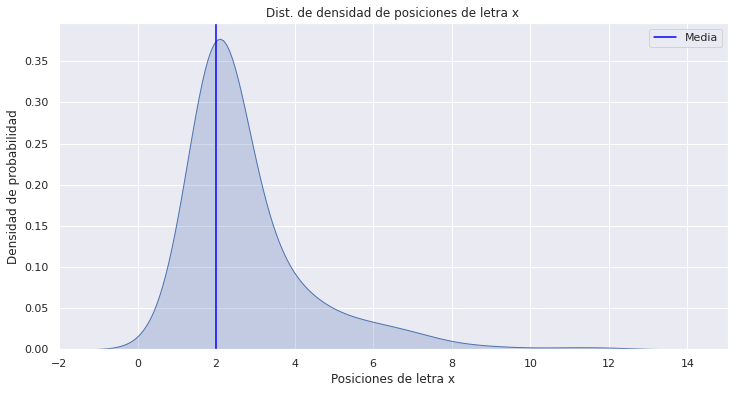


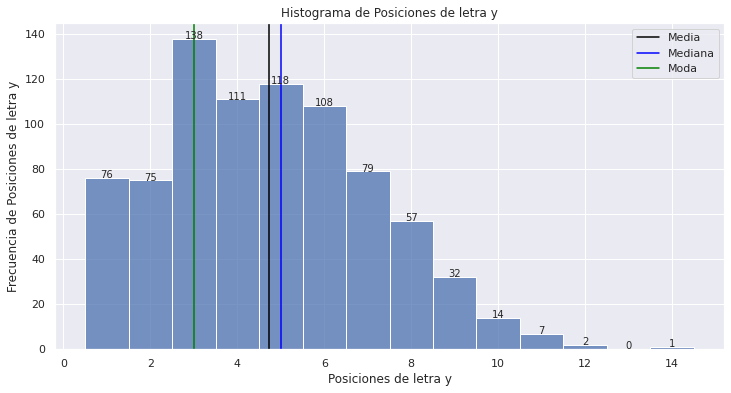


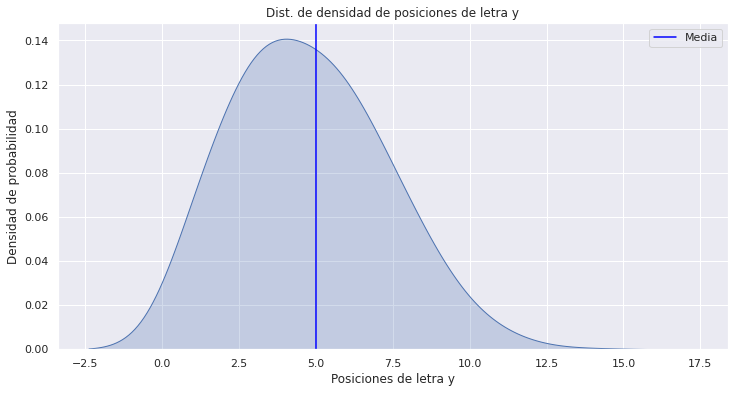


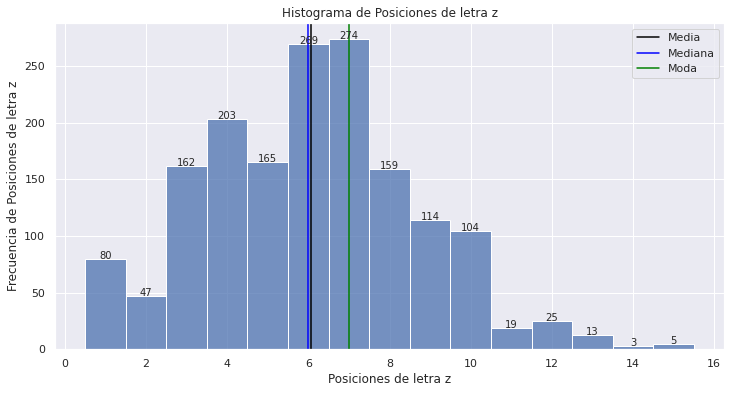


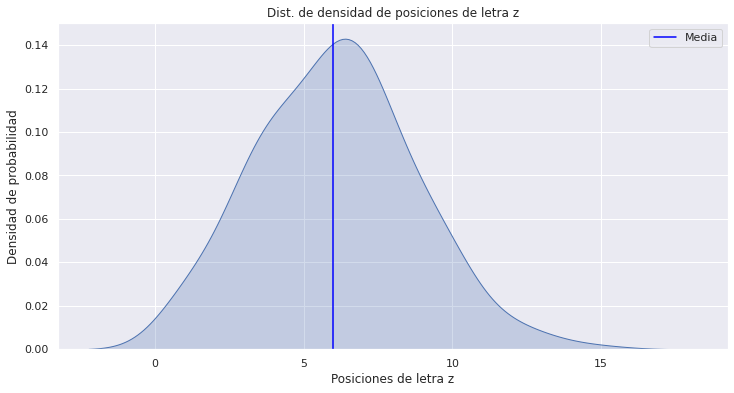


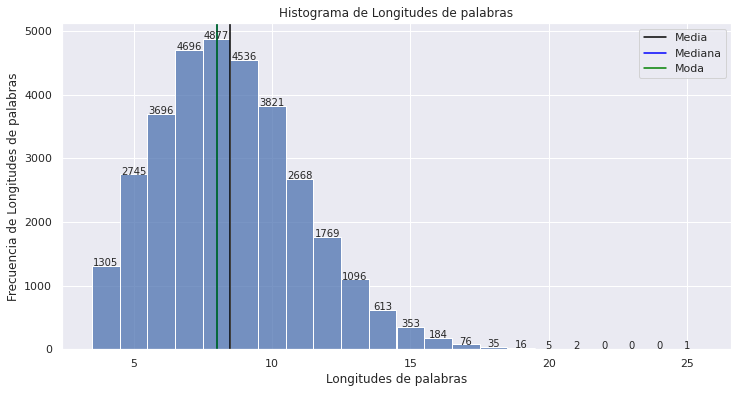


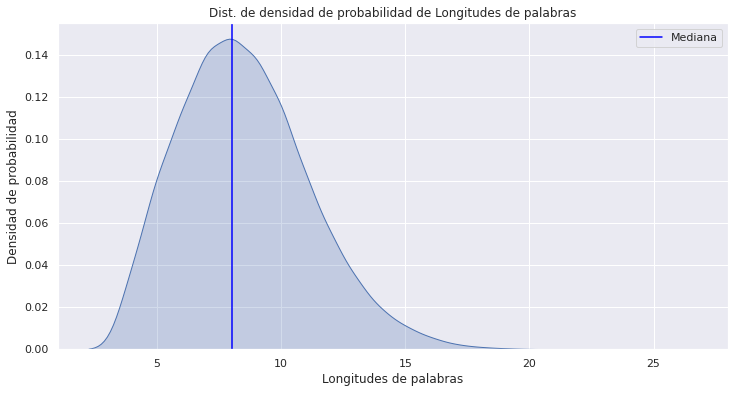




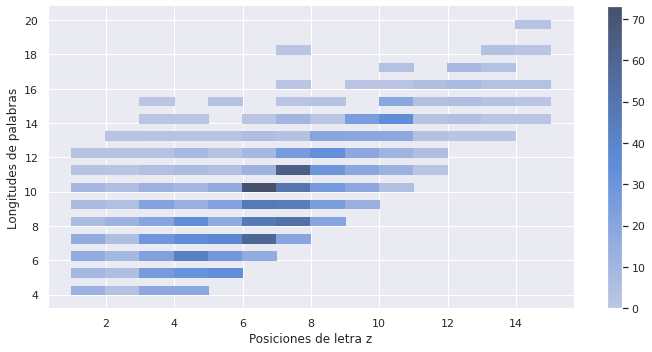
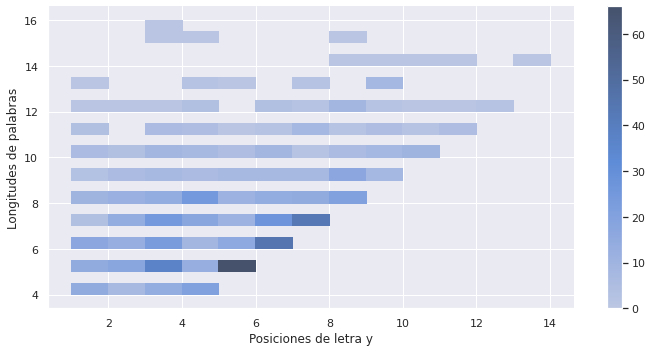
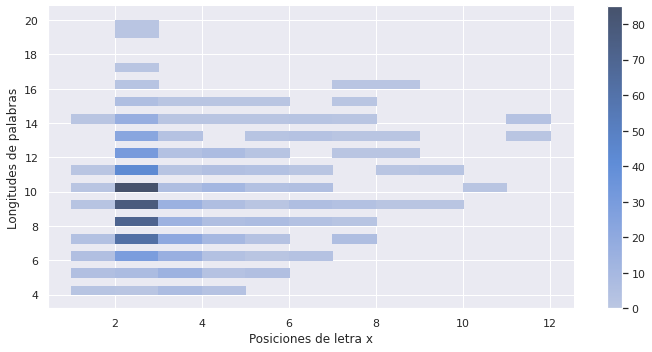
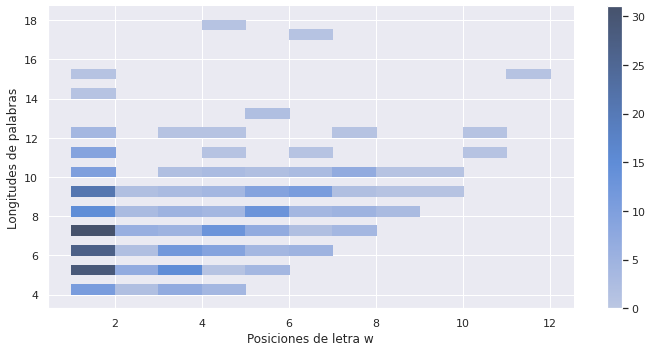
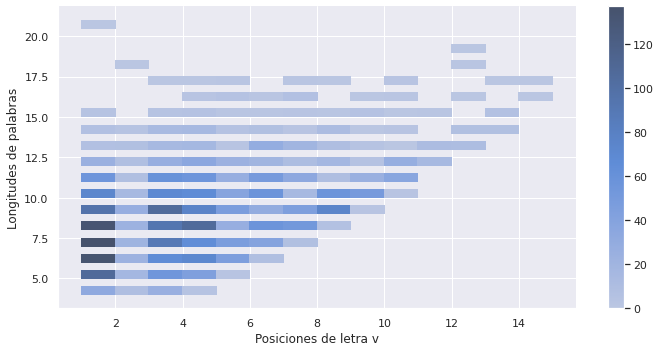
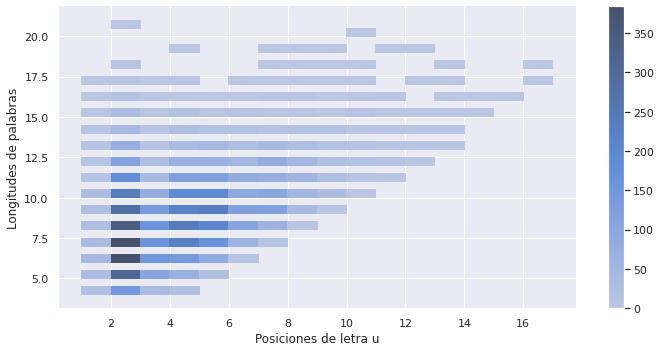
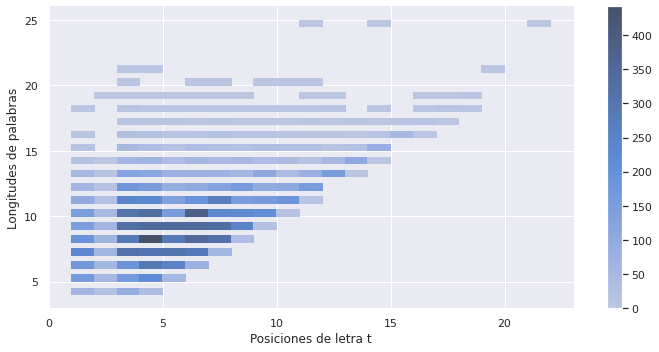
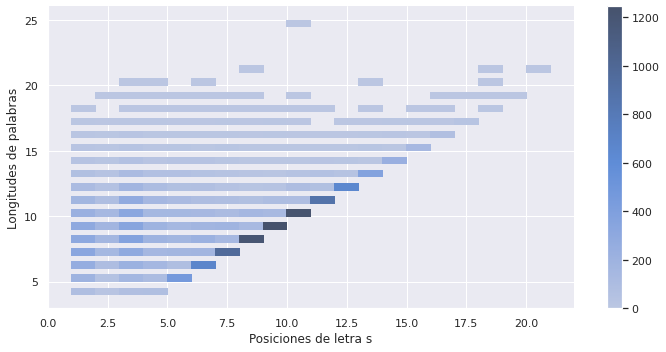
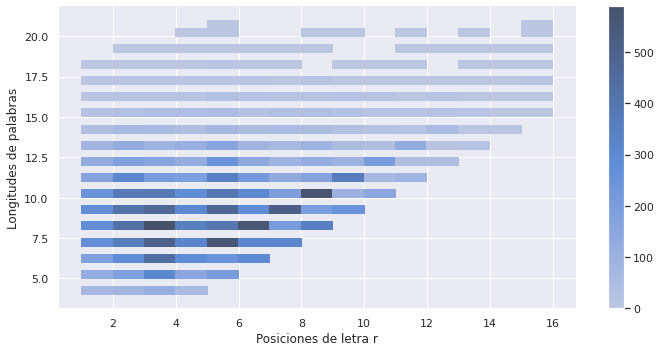
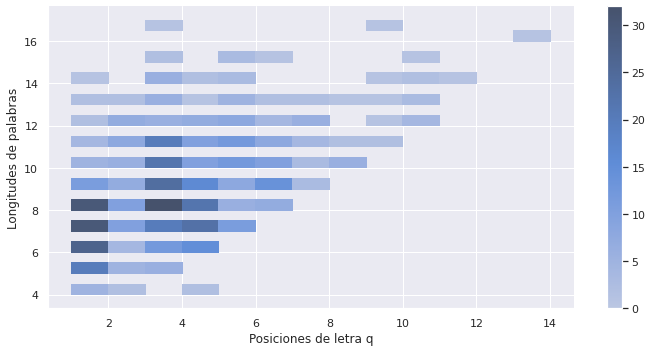
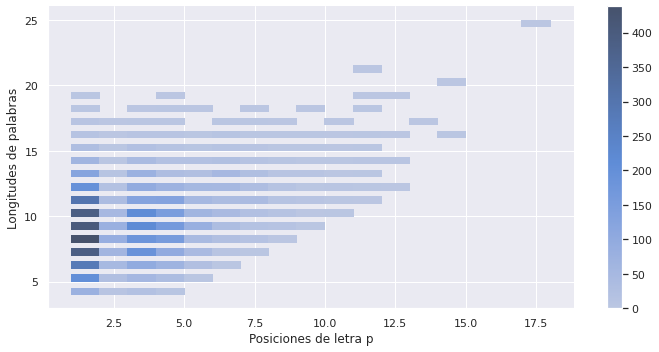
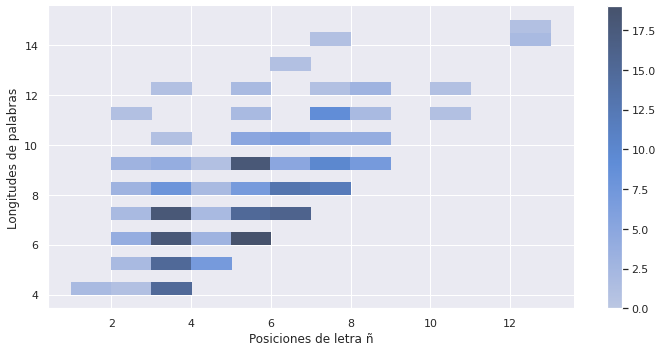
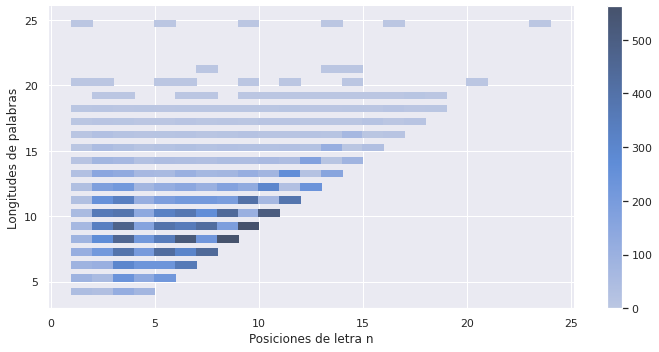
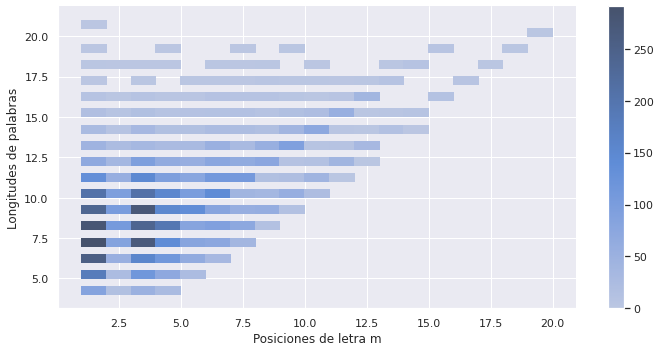
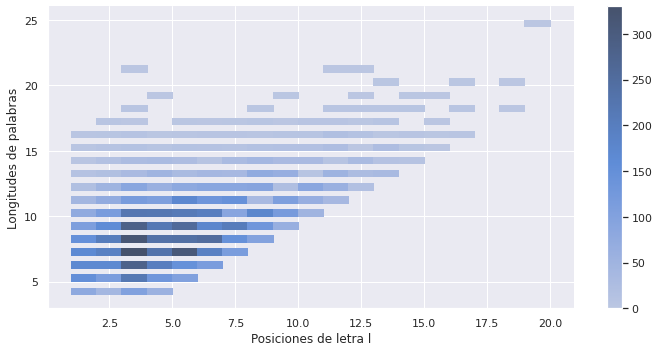
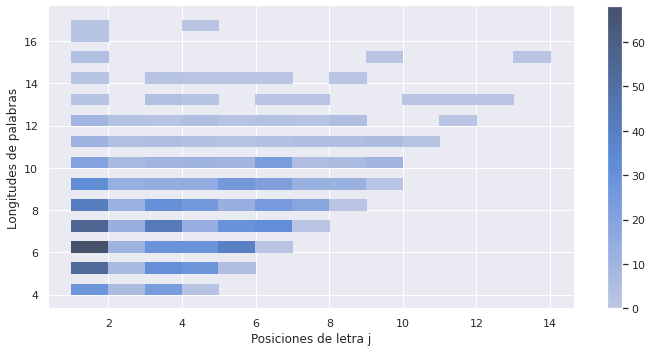
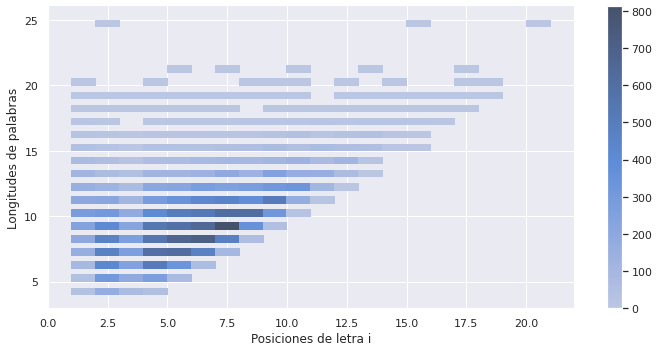
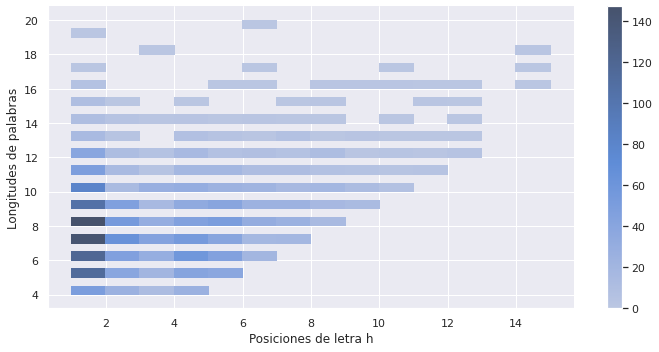
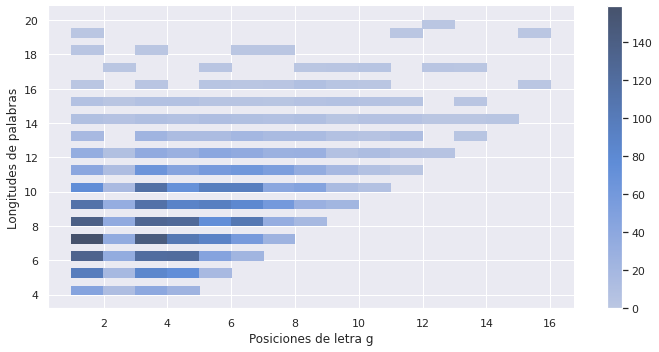
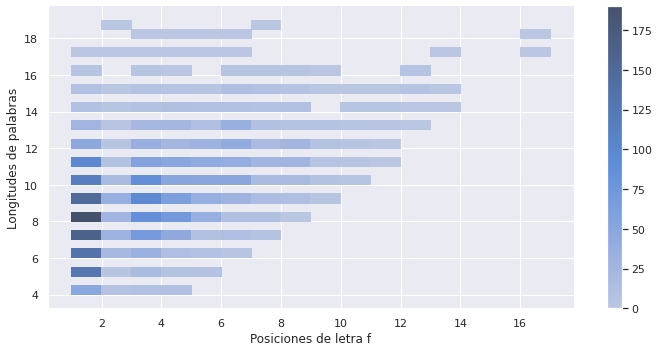
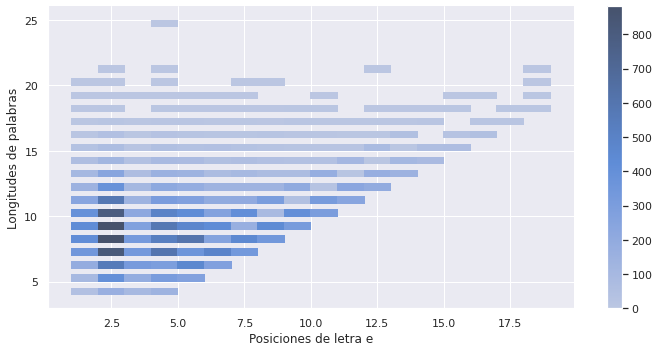
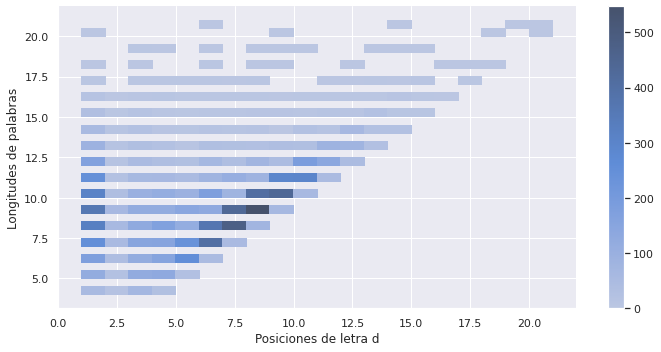
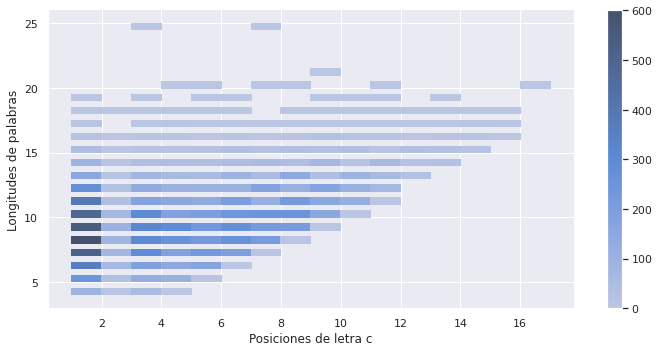
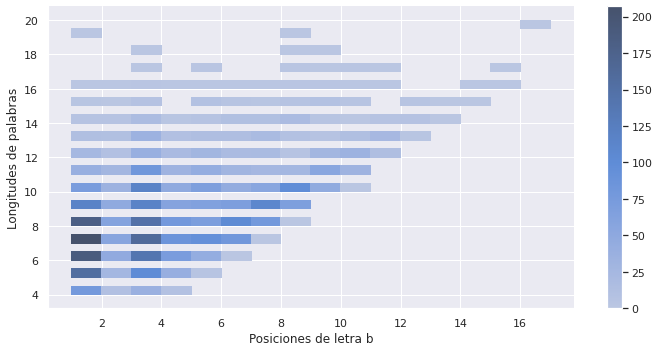
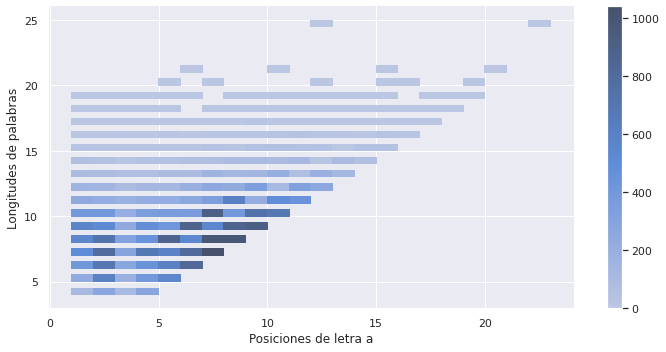








Por último, se realizaron gráficas que evidencian el comportamiento de la posición de las letras del alfabeto, respecto a la longitud de las palabras:



**Conclusiones:**

* Generalmente se percibieron diferencias en las frecuencias, posiciones, y las relaciones entre estas variables, para las letras del alfabeto español
* Se identificaron relaciones lineales positivas, y fuertes entre las cantidades de vocales y consonantes con la longitud de las palabras. Indicativo de la buena calidad de los datos.
* Mapas de densidad entre posiciones, frecuencias y longitudes de palabras permite evidenciar visualmente posiciones más frecuentes para letras y longitudes. Estos patrones son generalmente únicos por letra.
* Los hallazgos anteriores llevan a concebir una estrategia básica y eficiente de predicción de vocablos en el juego Wordle: es posible determinar qué letra es más probable que se encuentre en posiciones desconocidas, dada una longitud. (Recordar distribuciones de pos./Longitud)
* Se pueden calificar las palabras potenciales de acuerdo a cuáles concuerden más con las letras más probables en posiciones desconocidas.