Universidad Francisco Marroquín

Data Wrangling

19/11/2019

Censo Guatemala

Proyecto Final

Kirsten Jenatz - 20170113
Melany Donis - 20170474

Introducción

En el año 2018, el Instituto Nacional de Estadística (INE) realizó el Censo Nacional de Población y el Censo Nacional de Vivienda a lo largo de todo el perímetro del territorio guatemalteco, entre los meses de julio y agosto. El reporte muestra información sobre las características de la población guatemalteca, su sexo, edad, educación, pueblo e idioma, tecnologías de la comunicación, acceso a servicios, composición de hogares, estructura de vivienda a nivel departamental y municipal. Una limpieza y manipulación de estos datos es necesario para el análisis de los resultados. Los resultados se comprenderán de una manera gráfica para facilitar la exploración de su significado y de la situación actual de la población guatemalteca y las circunstancias en las que viven.

Metodología

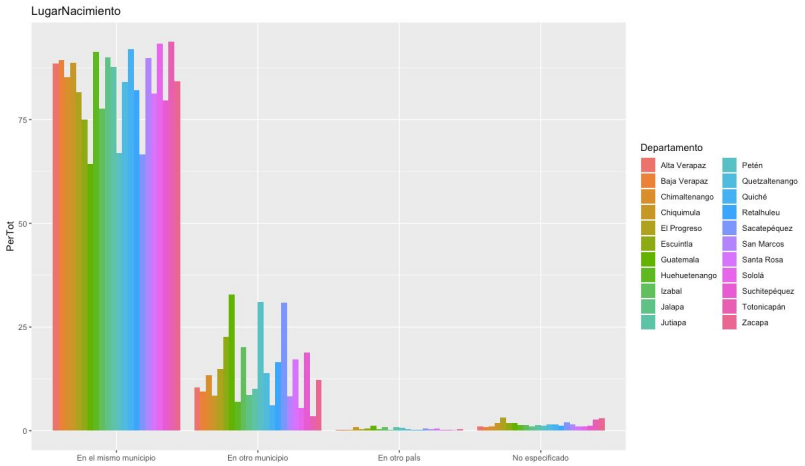
Para organizar los datos con los que se realizará el análisis, Tidy es la herramienta adecuada para generar un dataframe fácil de manipular. Al convertir los datos en tidy, cada fila se vuelve una observación, cada columna es una variable y cada entrada de la columna es un valor. Para esto, se utiliza el paquete tidyr de r que contiene las funciones de gather (crea pares clave-valor para asegurarse que todas las columnas sean variables), spread (separa filas que realmente son variables a columnas), separate (separa una columna con mucha información en varias) y unite (une múltiples columnas en una sola). En el caso de los datos del censo, primero se cambiará el nombre de las columnas, uniendo los 2 encabezados como una variable. Luego, se aplicará la función de gather a cada una y se separarán los nombres para poder diferenciar entre cada sub tema. Ya que son tantas tablas, es más conveniente mantener los datos así y unirlas, para luego poder seleccionar qué tema queremos investigar, y luego qué subtemas se necesitan.

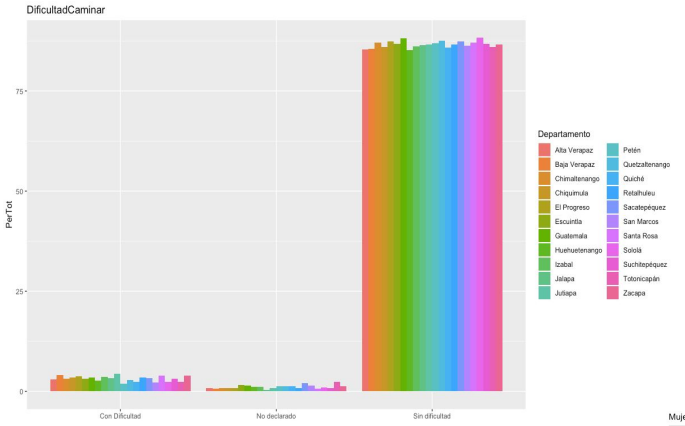
Exploración (EDA)

Gráficas

La mayoría de las comparaciones se hicieron en base a porcentajes porque debido a que Guatemala tiene una población mucho mayor a las de los demás departamentos los datos podrían estar sesgados a la hora de analizarlos.

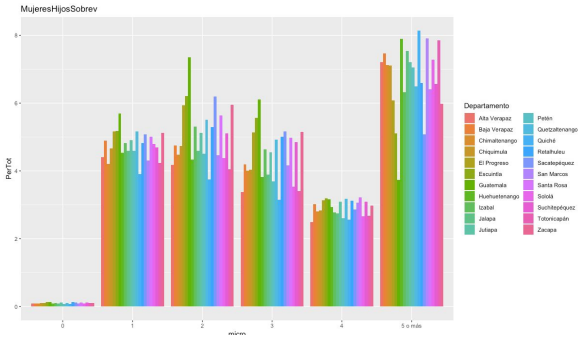
Si vemos que porcentaje de las personas de cada departamento realmente nacieron ahí, podemos observar que los tres departamentos a donde más se muda la gente es Guatemala, Petén y Sacatepéquez. El departamento donde más personas que nacieron ahí se quedan a vivir es Totonicapan.

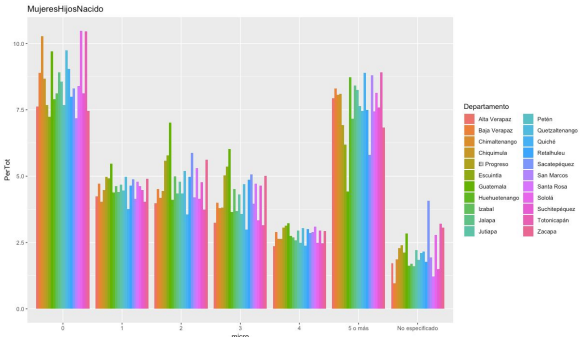


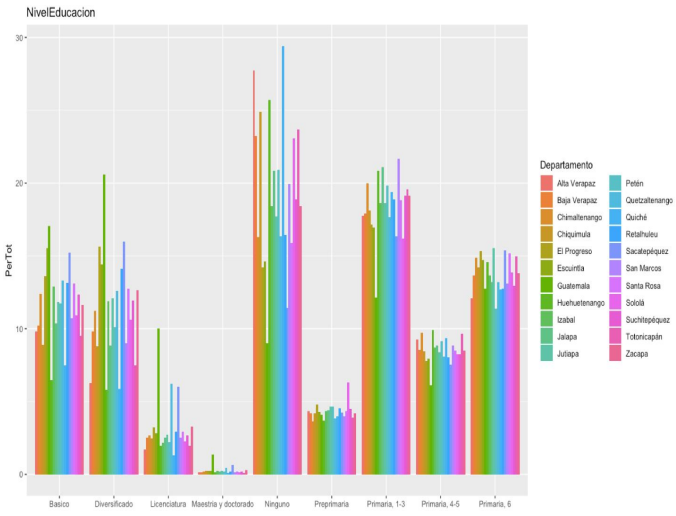


En general, en todos los departamentos más del 80 por ciento de las personas no tienen ninguna dificultad para caminar, oír, ver, recordar, etc. Las variaciones entre estos porcentajes son muy pequeños, con ningún departamento sobresaliente. Quizá lo interesante sea ver que en totonicapan un mayor porcentaje de las personas prefiere no responder si tienen alguna dificultad, manteniéndose para básicamente todas.

Otro dato interesante está basado en la cantidad de hijos sobrevivientes en los departamentos. Se puede notar que Guatemala tiene el mayor índice de supervivencia de 2 hijos, pero después de esto este porcentaje va disminuyendo. Esto se debe a que las personas en Guatemala tienen menos hijos que en otros departamentos, como se puede ver en la siguiente gráfica. Aunque es de los departamentos con porcentaje de un hijo nacido más altos, es el más bajo de 5 hijos o más. También se puede observar que varias personas decidieron no contestar la pregunta, por lo que concluimos que este puede ser un tema bastante sensible para las familias que perdieron algún hijo.

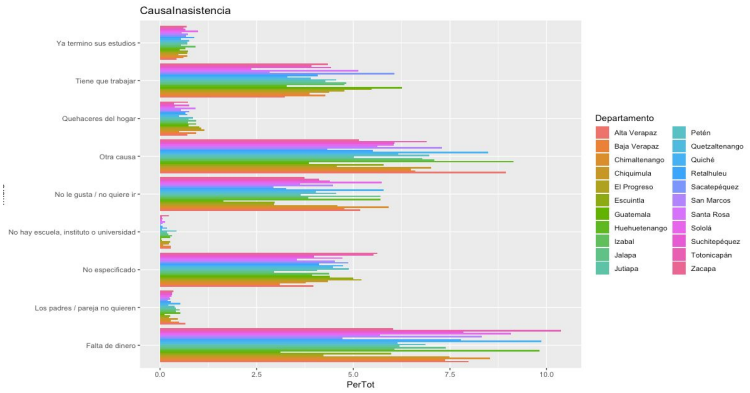


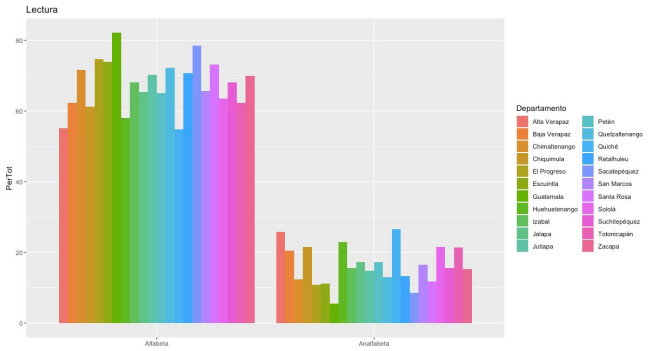




Como se puede observar en la siguiente gráfica, el porcentaje de personas que pasa por cada grupo de años escolar va disminuyendo. En varios departamentos más del 20% de la población no recibe ninguna educación. En todos los departamentos un pequeño porcentaje recibe solamente preprimaria, más un alto porcentaje deja el colegio entre los primeros tres años. Luego, si seguimos observando, se puede ver como el porcentaje de personas que recibe más educación va disminuyendo, pero por un menor porcentaje en guatemala, que claramente resalta en ambos licenciatura y maestría/doctorado.

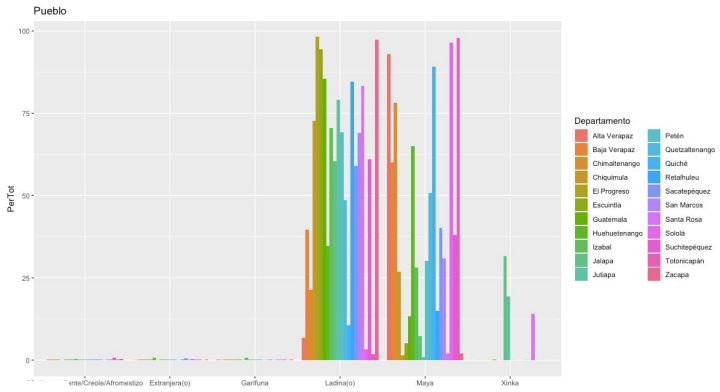
La siguiente gráfica demuestra que la mayor causa de inasistencia en la mayoría de los departamentos es la falta de dinero, no el hecho que tienen que trabajar. Sería interesante averiguar los precios que se deben pagar para ir a la escuela en estos departamentos, ya que si fuera gratis no debería de ser un factor la falta de dinero como tal.

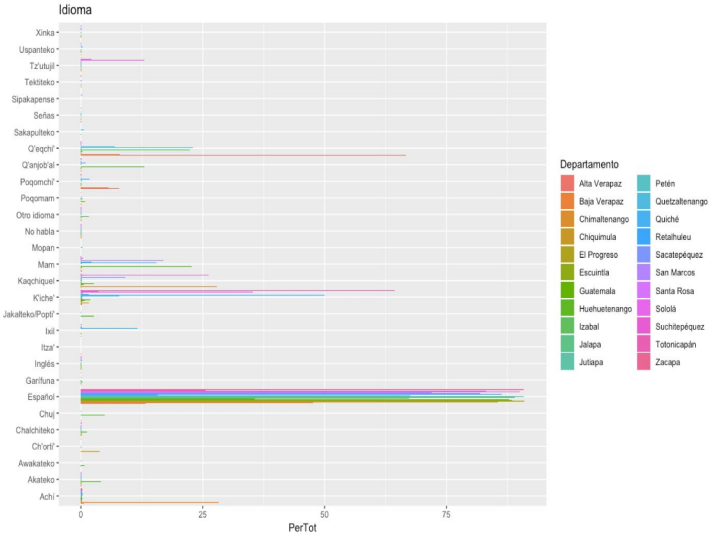




Un gran porcentaje de la población de los departamentos de Guatemala es analfabeta, con el menor porcentaje en el departamento de Guatemala. Por otro lado, el departamento con mayor porcentaje de personas analfabetas es Retalhuleu.

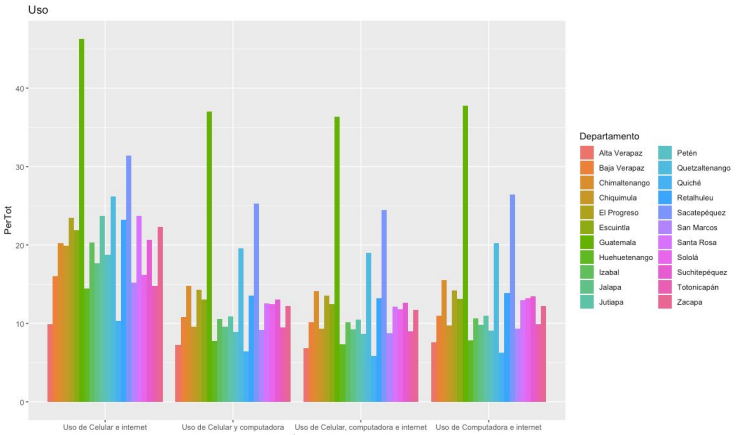
La mayoría de la población Guatemalteca se identifica como Ladino o Maya. Solamente en Santa Rosa, Jalapa y Jutiapa hay un porcentaje relativamente alto de personas que se identifican como Xinka. En Sololá y Totonicapán la mayoría de las personas se identifican como Mayas, y solamente un pequeño porcentaje se identifican como Ladinos.

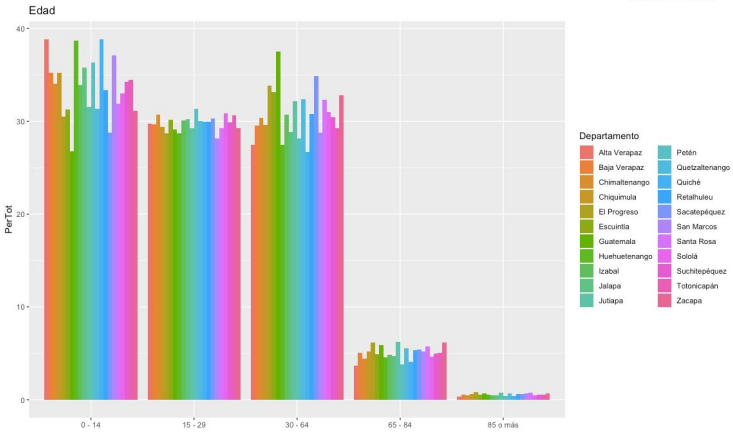




Como era de esperarse, varias personas hablan español en todos los departamentos. Solamente algunos idiomas son hablados por la mayoría de las personas en departamentos específicos; por ejemplo, en una de las verapaces habla más del 50% de las personas el idioma Q'eqchi'.

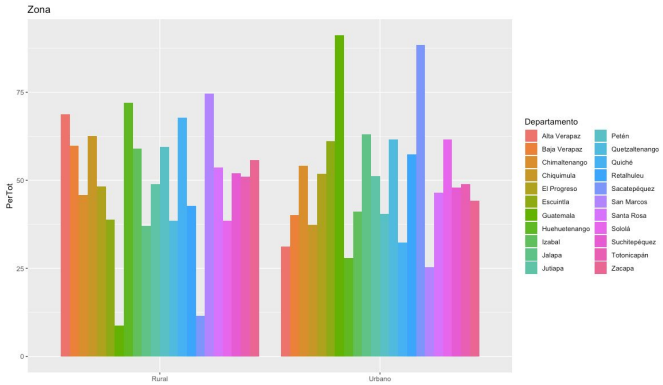
En la gráfica de la derecha se puede ver una gran diferencia entre el porcentaje de personas que utilizan ya sea computadora, celular o internet en Guatemala con los otros departamentos. El departamento donde la segunda mayor cantidad de personas utilizan estos dispositivos es Sacatepéquez. Esto asumimos que es porque varias personas viven en la Antigua pero se mantienen más en Guatemala.

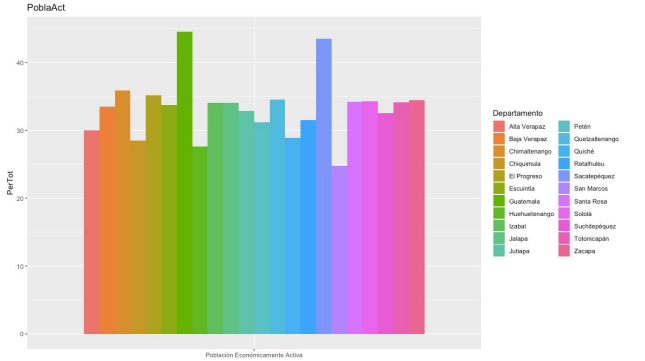


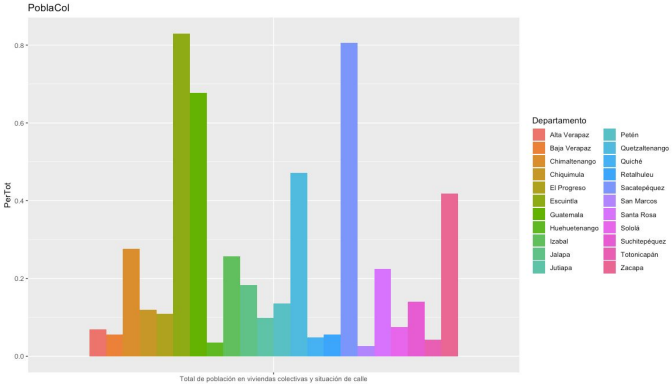


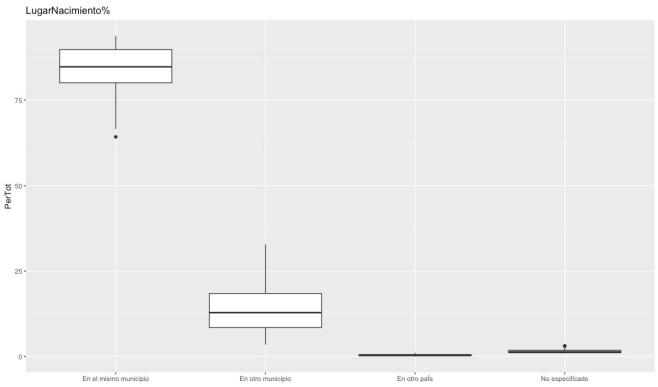
No se nota una diferencia extrema entre las edades de las personas de cada departamento, excepto que Guatemala tiene considerablemente menos niños y considerablemente más personas mayores. Esto muestra que las personas en este departamento no tienen muchos hijos.

A pesar de que se puede ver que la mayoría de las personas en el departamento de Guatemala y en el departamento de Sacatepéquez viven en zonas urbanas, en la gráfica de abajo se puede ver que es aquí donde el mayor porcentaje de personas (menos del 1%) vive en viviendas colectivas o en la calle. Además, en San Marcos hay más personas viviendo en zonas rurales pero menos en viviendas colectivas o en la calle. También es importante notar que un mayor porcentaje de la población está activa en Guatemala y Sacatepéquez que en los demás.



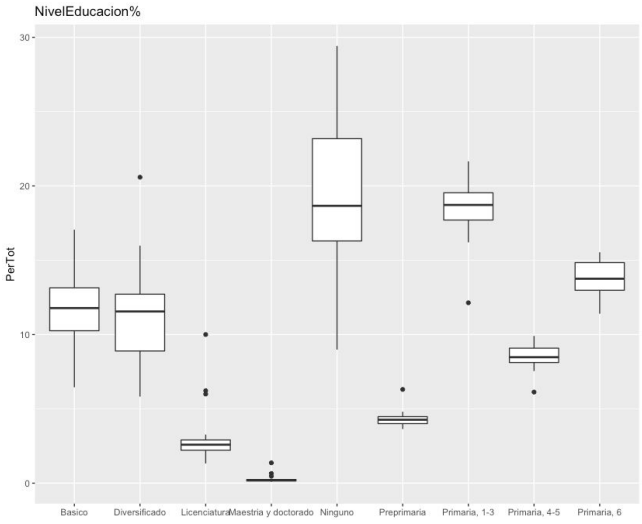


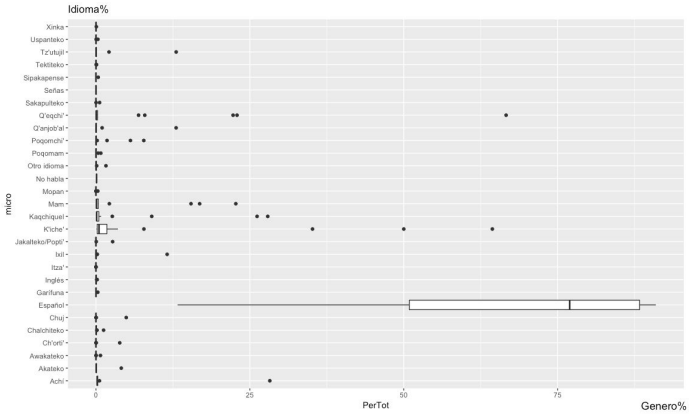




Si comparamos estos resultados con la primera gráfica que analizamos podemos asumir que el dato de Guatemala está muy por debajo de la media y podría ser considerado como un outlier. Sin embargo, no consideramos que está tan por debajo que pueda afectar los resultados, ya que no se encuentra tan lejos del cuartil inferior.

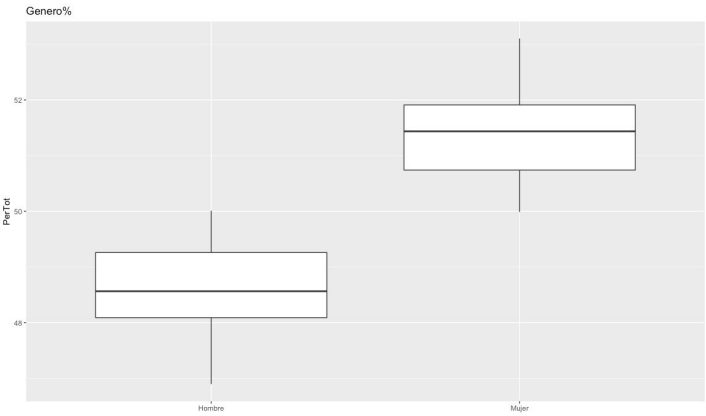
Si observamos los boxplots del nivel de educación, podemos ver que si existen varios datos atípicos. Sería interesante averiguar qué departamentos son estos datos, aunque podemos adivinar que la mayoría de los que están hacia arriba son en Guatemala o Sacatepequez ya que es aquí donde la mayoría de las personas reciben más educación. Aquí podemos volver a recalcar que un porcentaje bastante alto de todos los departamentos no han recibido ninguna educación, con la media cerca del 20%.



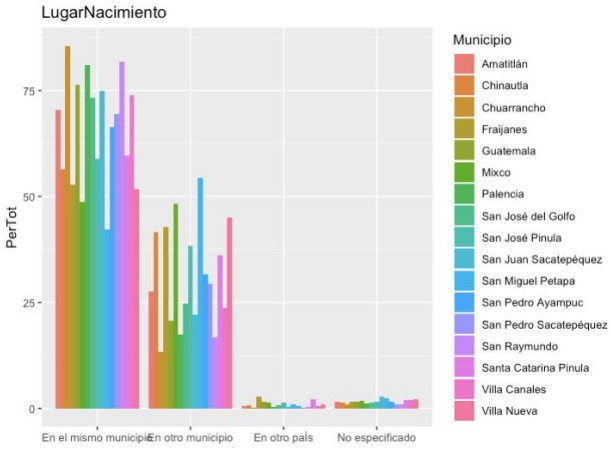


En la gráfica a la izquierda se puede ver que el único idioma que los departamentos tienen en común es el español. Por eso existen tantos datos atípicos en esta gráfica; la mayoría de los datos de otros idiomas serán muy bajos y afectarán la media.

Si vemos esta gráfica podemos ver que la distribución entre hombres y mujeres es bastante equitativa, pero hay un poco más mujeres que hombres. No parecen haber datos atípicos en esta parte, por lo que está distribución se mantiene en todos los departamentos.

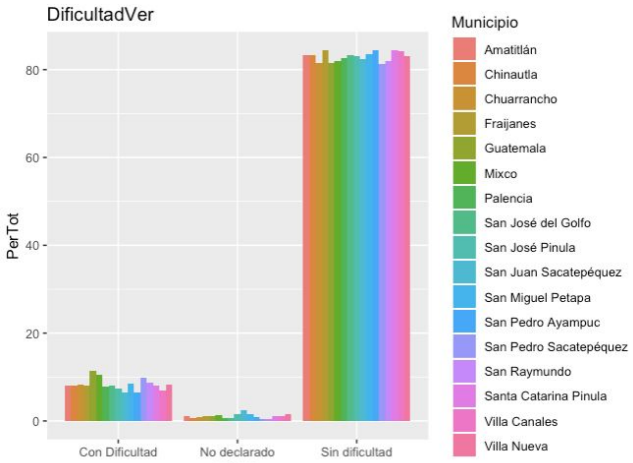


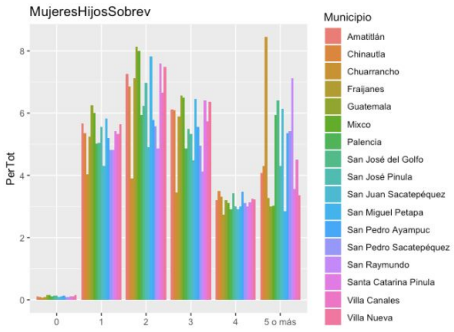
Para el análisis de los municipios se analizará los municipios que pertenecen al departamento de Guatemala, ya que en el análisis realizado previamente a los departamentos se demostró que éste es el departamento que representa un mayor porcentaje de los datos, debido a su gran población.

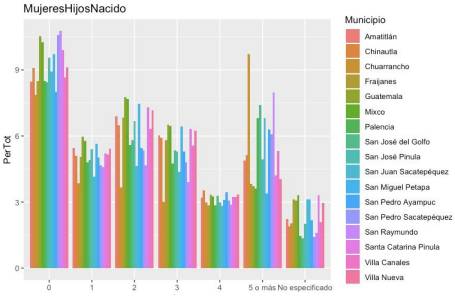


La gráfica indica que Fraijanes y San Raymundo son los municipios en donde la mayor cantidad de personas que nacen ahí, se quedan en ese municipio. El municipio en donde más personas se van a vivir es San Miguel Petapa, Mixco y Villa Nueva. Lo cual es interesante ya que en lugar de ser la ciudad, como se esperaría, las personas prefieren irse a vivir a los municipios a las afueras de la ciudad. También indica que no hay muchas personas que nacieron en otro país y viven en Guatemala.

Más del 80% de las personas en todos los municipios no tienen ninguna dificultad para ver. Sin embargo, casi el 10% tiene dificultades, esto se puede deber a que con dificultad las personas se refieren a que tienen la necesidad de usar lentes, no necesariamente si pueden ver o no. Esto sucede especialmente en los municipios de Guatemala y Mixco.

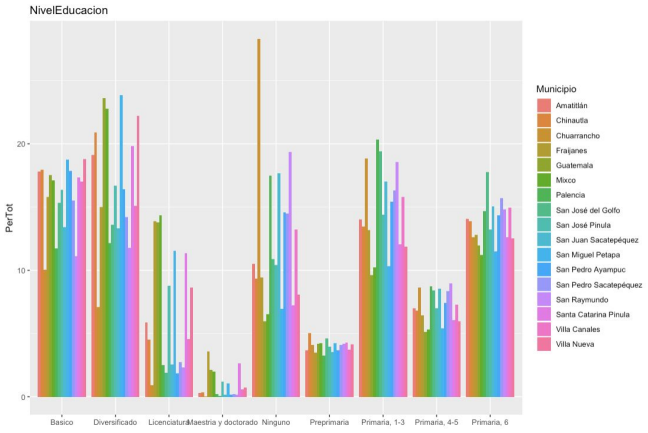


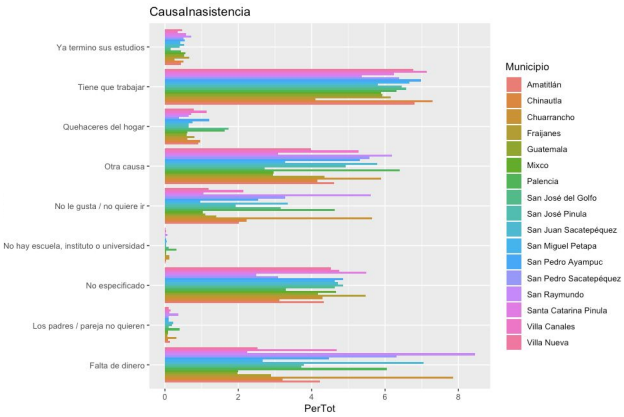




El municipio en donde más hijos se tienen es en chuarrancho y es por esta razón que en donde sobreviven más hijos es en ese municipio. Siguiente municipio en donde las personas tienen más de 5 hijos y que sobreviven es en San Raymundo. Fraijanes es el municipio en donde más personas han tenido 2 hijos y han sobrevivido. Además es evidente que existe una gran cantidad de mujeres en todos los municipios que aún no han tenido hijos. Sin embargo, esta cantidad no es tan grande si juntamos a las personas que tienen de 1 a más de 5 hijos. Es interesante notar que la cantidad de hijos más común en el departamento de Guatemala es dos, lo que podría ser contrario en otros municipios fuera de este departamento.

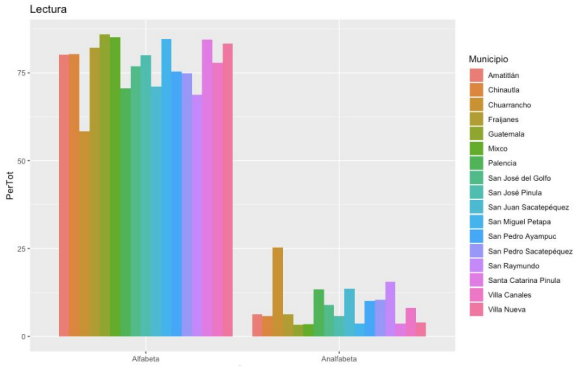
El municipio en donde la mayoría de las personas no tienen algún nivel de educación es Chuarrancho. Es evidente que la mayoría de los habitantes del departamento de Guatemala llegan como mínimo a graduarse de diversificado, pues muchas personas deciden empezar a trabajar en lugar de seguir estudiando en la universidad, especialmente en los municipios de la ciudad, Mixco, San Miguel Petapa y Villa Nueva. Además, muy pocas personas deciden seguir una maestría o doctorado y muy pocas reciben una educación preprimaria, ya que entran a estudiar muchas veces hasta primero primaria.

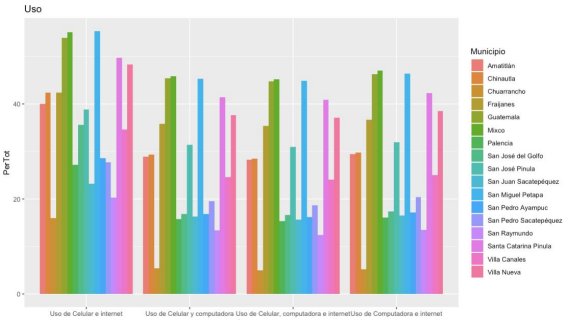




La causa principal por la que la mayoría de las personas no estudia es porque trabajan, especialmente en los municipios de Chinautla y Villa Canales. El municipio de Chuarrancho es en donde una mayor cantidad de personas no estudias por causa de que no tienen dinero, no les gusta o no especificaron la razón. Las personas de San Raymundo no estudian porque no les gusta o no quieren ir, lo cual demuestra que esa no es una prioridad para ellos.

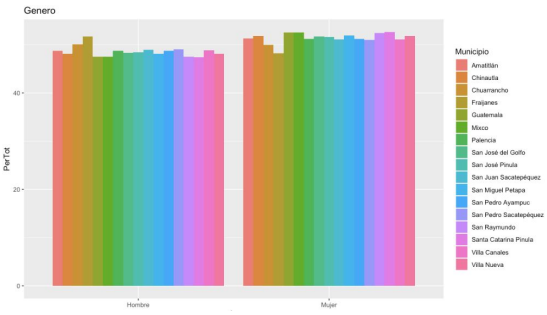
Más del 75% de las personas son alfabetas, siendo guatemala el municipio con mayor porcentaje. Por otro lado, en Chuarrancho es en donde existe un mayor porcentaje de personas analfabetas, lo que puede ser el resultado de lo que se analizó anteriormente, indicando que en ese municipio muchas personas no estudian y por lo tanto no tienen un título de algún nivel.

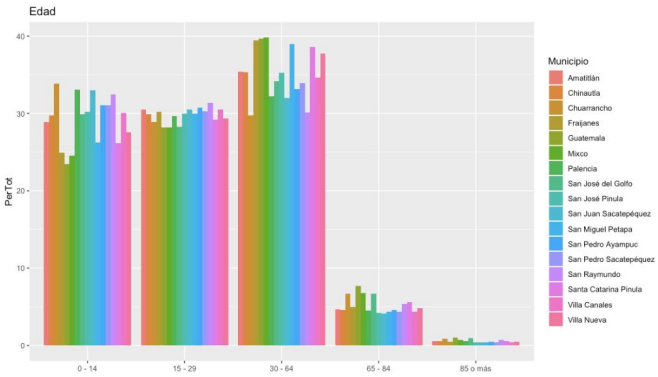




Es
evidente que en Guatemala, Mixco y San Miguel Petapa es donde se da el mayor uso de celular, internet y computadora. Sin embargo, el uso del celular e internet es el que tiene mayor porcentaje, lo que se debe a que en Guatemala hay más líneas de teléfono que personas.

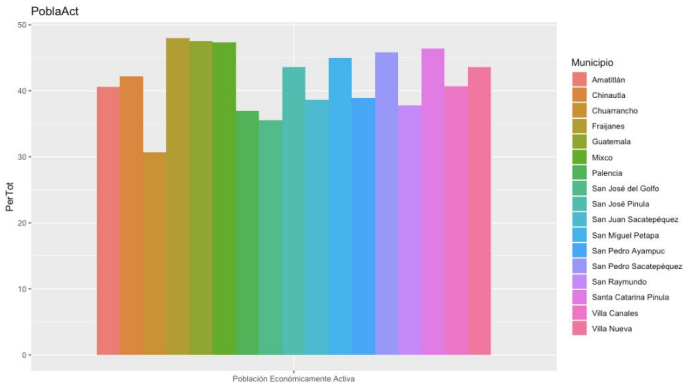
La gráfica demuestra que la población de Guatemala está constituida casi 50/50 entre hombres y mujeres. Sin embargo, sorprendentemente en la mayoría de los municipios existe un mayor porcentaje de mujeres que hombres, en promedio, aunque la diferencia sea menor al 5%. Aún asi, existe una mayor cantidad de hombres en Fraijanes y Chuarrancho.



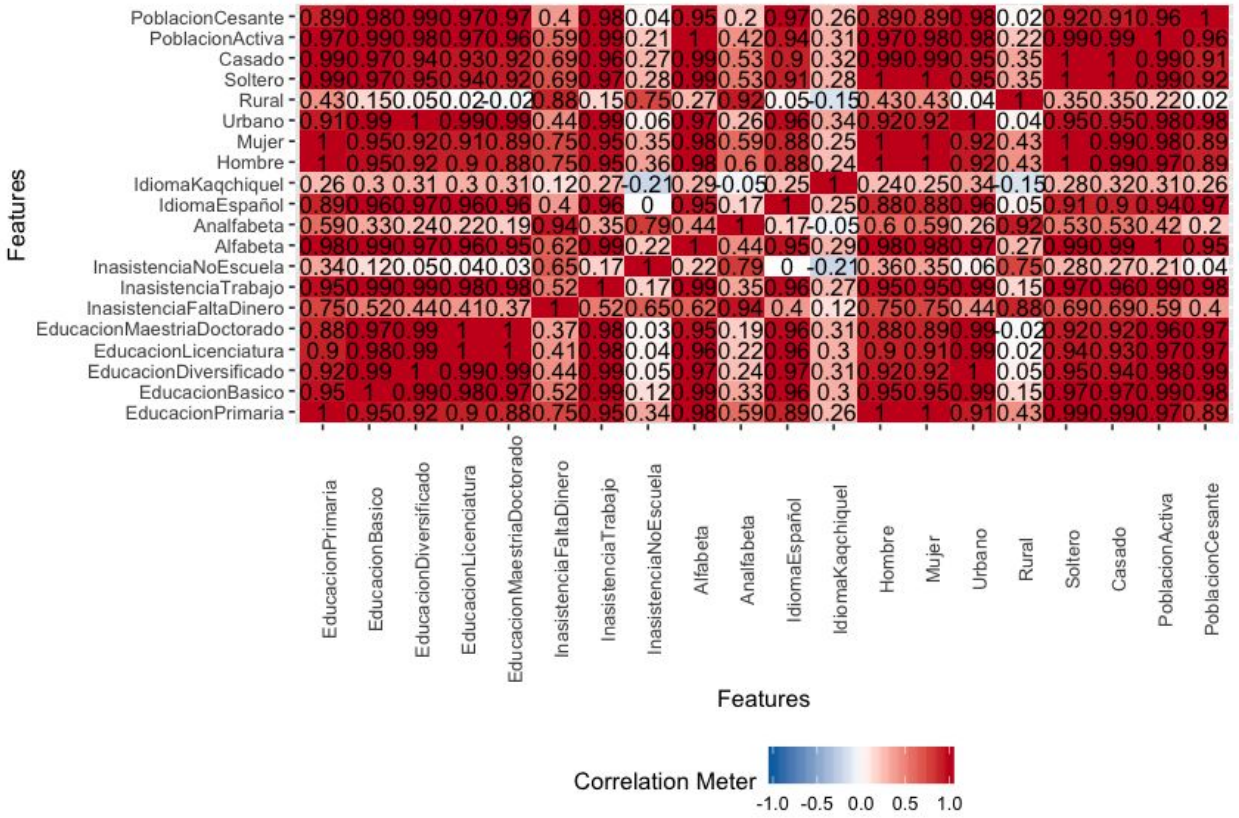


La mayor cantidad de habitantes son menores de 64 años, especialmente en los municipios de Fraijanes, Guatemala y Mixco. El mayor porcentaje se mantiene entre las edades de 30-64, sin embargo la población joven es muy grande comparado con otros países desarrollados.

Como era de esperarse según lo visto anteriormente, Chuarrancho es el municipio en donde el porcentaje de su población económicamente activa es muy bajo, debido a que sin educación es muy difícil conseguir empleo. Por otro lado, los municipios de Fraijanes, la ciudad, Mixco y Santa Catalina Pinula es en donde las personas aportan más a la economía del país.



Correlación de variables

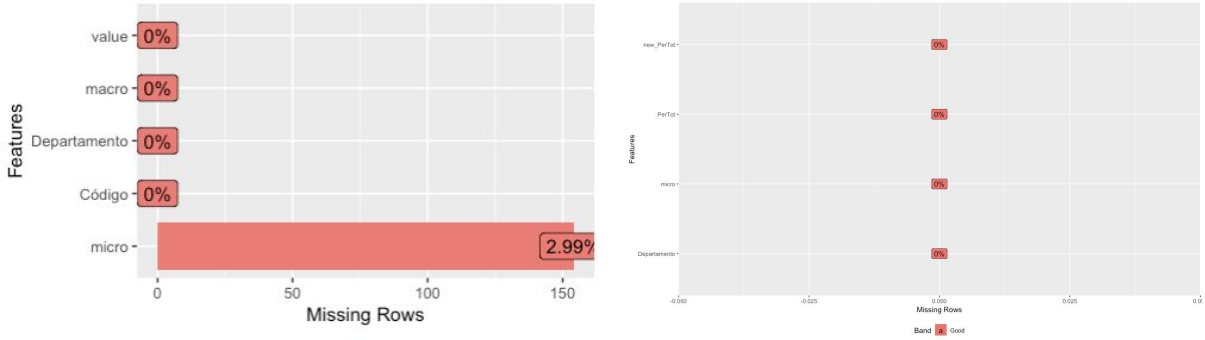


La gráfica de correlación realizada sobre las variables más importantes y con cantidades significativas indica que la mayoría de estas variables están correlacionadas entre sí, pues mientras más rojo sea su color, mayor es la correlación que comparten. Se puede observar que los diferentes grados de educación están muy relacionados, ya que es mucho más probable que las personas que estudian educación primaria terminan estudiando hasta una licenciatura y en el mejor de los casos una maestría o doctorado. La correlación de este último va aumentando mientras más alto el nivel educativo. La causa de la inasistencia de las personas porque trabajan está correlacionada con todos los niveles de educación y con la población económicamente activa, lo que significa que hasta los niños que estudian primaria faltan debido a que trabajan. Además, es evidente que la inasistencia por falta de dinero es menos común mientras mayor el nivel de educación, lo que es muy probable debido a que si llegan a estudiar una maestría es porque tiene el dinero necesario como para que esa sea la causa de su inasistencia. Sin embargo, la inasistencia por falta de escuelas es el menos correlacionado con los niveles de educación. El alfabetismo está altamente correlacionado con todos los niveles de educación, pues si estudian es elemental que sepan leer, además, los que hablan en español son los más comunes en estos. Por consiguiente, es razonable que el analfabetismo y los que hablan un idioma diferente al español están menos correlacionados. La zona urbana está más relacionado con todos los niveles ya que la educación es más accesible en esa área, en contraste con la zona rural. Además las personas económicamente activas están correlacionadas con los niveles de educación, pues es más probable conseguir un trabajo con un título profesional. El analfabetismo está poco relacionado con las variables, a excepción del área rural, pues es evidente la poca oferta de una digna educación.

Feature engineering

Ya que no se está creando ningún modelo predictivo, realmente no aplica hacer mucho feature engineering. Por lo tanto, solo nos enfocaremos en que ocurrió al momento de intentar transformar los nombres de las columnas y un intento de cambiar cualquier outlier que salió durante el exploratory data analysis.

Missing Values: En realidad no existen missing values en la data que recopilamos. Lo único que ocurrió es que cuando se hizo el gather para crear las columnas de macro, micro y value, faltaron algunos valores en la columna de micro. No obstante, esto fue fácil de componer ya que solo se le imputo el mismo nombre que la columna de macro, porque estas no tenian ningun sub-tema y no tenían que ser agrupadas.



Outliers: No es razonable excluir los datos atípicos de nuestro análisis, pues estos representan departamentos y municipios que no pueden ser omitidos o cambiados sin sesgar o alterar los resultados. No solo se estaría borrando un dato, sino información de un departamento o municipio entero del país, el cual representa datos de miles de personas. Además, si uno quiere averiguar sobre la situación de guatemala lo que le interesa son los outliers, por lo que no deben de ser borrados o afectados de ninguna manera. Creemos que en este tipo de investigación, lo que se debería de hacer es buscar formas de mejorar los datos que estén muy por debajo de lo que deberían estar, o buscar razones por las cuales algunos otros son más altos comparados con los demás. Así, los programas e instituciones gubernamentales pueden crear proyectos para solucionar estos problemas.

Hallazgos

Los resultados que se obtuvieron de los datos recopilados en el censo de población del año 2018 fueron muy interesantes. Primero, el censo mostró que 14.9 millones de personas viven en Guatemala. De estos, 7.22 millones corresponde a hombres, es decir el 48.5 % del total, mientras que 7.67 millones son mujeres; el 51.5% del total. Es interesante notar que, aunque la diferencia entre ambos porcentaje no es significativamente grande, hay más mujeres guatemaltecas que hombres. La mayor cantidad de esta población se encuentra en los departamentos de Guatemala, Alta Verapaz, Huehuetenango, y San Marcos, debido a éstos son los departamentos más desarrollados con los que cuenta el país. Por otro lado, de estos departamentos solamente la población de Guatemala y San Marcos aporta más a la economía del país en comparación a los demás, casi 10% más. Según las respuestas de los guatemaltecos, el 56% se consideran como ladinos, mientras que el 41.7% se identifica como maya, y el resto se identifican como xinca,

afrodescendiente, garífuna o extranjero. Esto significa que aún cuando la población indígena representa casi la mitad de toda la población, la mayor parte de la población guatemalteca es ladina, siendo el español el idioma más común entre ella.

Aunque el nivel de alfabetismo del país es de 81.5%, con el mayor porcentaje viviendo en Guatemala, existe un porcentaje altamente preocupante de la población, casi un 30%, sin ningún nivel educativo en los otros departamentos. Esto denota que la mayor parte de la población entera carece de una educación adecuada, lo cual afecta su calidad de vida y su aporte a la economía. Por otro lado, Chuarrancho se destaca como el municipio con mayor porcentaje de personas sin educación, analfabetas y con menos personas económicamente activas dentro del departamento de Guatemala, ya que éste es el departamento con mejores índices.

Conclusiones

Buscamos un poco acerca de las personas registradas en RENAP. En diciembre de 2018 habían 21.6 millones de guatemaltecos registrados. Como se puede notar, este número es un tercio más de lo que realmente se recolectó en el censo realizado el mismo año, y muchas personas han comenzado a dudar del trabajo que RENAP realiza. Sin embargo, se debe tomar en consideración que RENAP no cuenta con que algunas de la personas registradas no viven en el país. Tampoco se toma en cuenta el hecho que en el censo no se llegó al cien por ciento de la población. Si asumimos que el gobierno no puede llegar al 5% de los habitantes por causas externas (condominios no permiten el ingreso o aldeas a las que no se puede llegar con facilidad), el número de personas en Guatemala aumentaría a 15.6 millones de personas en el país. También pueden haber casos donde hay personas difuntas que no han sido registradas en RENAP, aumentando la cifra de personas registradas momentáneamente. Si no se registra la muerte de una persona, RENAP no comienza a investigar hasta que se vea que alguien llegue a una edad mayor, como 110 años, por lo que parte de esos 21 millones de personas pueden no estar viviendo actualmente.

En conclusión, los datos del censo permiten conocer la situación actual del país, determinando cuántas personas hay en guatemala, que hacen y qué necesidades tienen. Esto puede ayudar a orientar las acciones de las instituciones y programas gubernamentales, con el objetivo de resolver correctamente los problemas que se establecieron con este análisis. Si basaran las decisiones de estos programas e instituciones solamente en los datos de registros de RENAP, no se estaría estimando correctamente la población guatemalteca y muchos proyectos podrían fracasar.

Recomendaciones y siguientes pasos

Recomendamos analizar con más profundidad los demás municipios, ya que nosotras solo nos enfocamos en los municipios del departamento de Guatemala por la importancia que representan en nuestro análisis. Quizá lo más razonable sea analizar los municipios de los departamentos que aparezcan como datos atípicos, para poder ver si es solo un municipio el que afecta el resultado o si es todo el departamento el que se diferencia de los demás datos.

El siguiente paso para esta investigación sería buscar datos de censos de años anteriores. De está manera se podrá observar cómo han ido cambiando los datos a lo largo del tiempo y cómo se podrían llegar a ver en el futuro. Así, se puede comenzar a analizar las causas de estos cambios y quizá hasta se podría

intentar forzar el crecimiento o la mejora de la situación de varias familias en base a estas variables. También sería interesante de alguna manera conseguir información sobre la población migrante, ya que varias personas Guatemaltecas viven en el extranjero. Se podrían analizar las causas por las que se van, por ejemplo; si su situación de vida en Guatemala era muy mala y decidieron irse o están trabajando/estudiando fuera temporalmente.

Anexos

En la siguiente pagina se pueden encontrar los datos que han sido registrados en RENAP y que fueron analizados en nuestras conclusiones:
https://lahora.gt/renap-registra-casi-22-millones-de-guatemaltecos-inscritos/

Variables: Se creó un listado de variables para cada tabla que determinaba el nombre de la primera columna, la cual se volvería más adelante la columna macro. Estos nombres se cambiaban al importar los datos.

Funciones: A continuación explicaremos las funciones que utilizamos para analizar los datos:

- *Read Function:* Se creó una función que importa todas las tablas. Para los departamentos, se limitó las filas de la número 9 a la 32, para importar únicamente la tabla con los datos que se necesitaban de los 22 departamentos. En el caso de los municipios se limitó las filas de la número 9 a la 350 para importar los datos de los 340 municipios.
- *Names Function:* Esta función se utilizó para unir los dos encabezado de cada tabla en uno solo, separadores con "_". Esto solo lo hacia si la columna tenía dos, de lo contrario dejaba el único que tenía.
- *Gather Function:* Con ella, se volvió tidy los datos, uniendo las columnas como un macro y su valor, para luego separar el macro en 2 los dos encabezados que se habían unido anteriormente.
- *Histogram Function:* Se creó una función que realizaba un histograma para cada elemento de una lista creada de la columna macro.
- *Boxplot Function:* Se hizo lo mismo que con los histogramas pero para crear boxplots.

Entregable: Los archivos dep.RData y mun.RData en el folder son las listas que se crearon con cada variable de forma tidy como un environment en R.