Universidad Francisco Marroquín

Juan Carlos Girón

José Rolando Josué

21/11/2019

Contraste de la región Metropolitana de Guatemala con el resto del territorio nacional

Data Wrangling: Proyecto Final

Fausto Holcombe - 20170476

Guillermo Puente - 20170036

**Introducción**

El censo es el listado oficial de los habitantes de un país o estado, el censo se usa para poder tener registro de la cantidad de personas que habitan un país y también de diversos datos de dichas personas. En Guatemala en este año 2019 fue el “XII censo de población y VII censo de vivienda” para poder recaudar información de la población y de distintas condiciones en las que vive. El último censo que se llevó a cabo en Guatemala, el “XI censo de población y VI censo de vivienda”, fue en el año 2002. Son muchos años de diferencia para hacer el censo, 17 específicamente, considerando que un gobierno debería hacer un censo cada 5 años aproximadamente.

Los departamentos de Guatemala pueden ser separados por distintas regiones, estas tienden a estar separadas por el lugar en el que se encuentran dentro del territorio nacional. Las regiones son las siguientes:

* Metropolitana: Conformada únicamente por Guatemala.
* Petén: Conformada únicamente por Petén.
* Noroccidental: Conformada por Huehuetenango y Quiché.
* Suroccidental: Conformada por Sololá, Totonicapán, Quetzaltenango, Suchitepéquez, Retalhuleu y San Marcos.
* Central: Conformada por Sacatepéquez, Chimaltenango y Escuintla.
* Suroriental: Conformada por Santa Rosa, Jutiapa y Jalapa
* Nororiental: El Progreso, Izabal, Zacapa y Chiquimula.
* Norte: Alta Verapaz y Baja Verapaz.

En el reporte detallaremos el análisis que se llevó a cabo de la investigación usando métodos recibidos en clase, además de arreglar el formato de las tablas para que la búsqueda y extracción de información sea de mayor facilidad y comodidad. Se buscará encontrar información relevante e importante en las tablas para poder tener una percepción de la realidad que vivimos en estos momentos.

**Objetivos**

En el trabajo que se está llevando a cabo se busca cumplir con diferentes objetivos, tanto en el aparte de análisis como en la parte de limpieza de datos:

* Arreglar y modificar las tablas para facilidad de análisis.
* Usar métodos aprendidos en clase para poder llevar a cabo dicho arreglo y así tener las tablas mejor relacionadas y de manera más ordenada.
* Encontrar información que pueda ser relevante y que pueda proveer un *insight* sobre la realidad de cada región.
* Visualizar y encontrar diferencias entre el desarrollo y estado de cada región en relación con la región metropolitana.

**Metodología**

Para el manejo y arreglo de las tablas se usaron distintos paquetes, por ejemplo, se usó “Readxl”, “ReadR” y “Zoo” para importar las tablas y además se usaron funciones que incluían algunas partes de dichos paquetes -y por supuesto funciones que vienen ya instaladas en R-. Esas funciones se crearon específicamente para solucionar problemas con el formato en el que las tablas venían descargadas, problemas como el espaciado y el hecho que algunas de ellas estaban agrupadas por categorías distintas con celdas que solamente se encontraban encima de algunas celdas y encima de otras había otras clasificaciones o podía no haber.

Ahora, se crearon otras funciones utilizando herramientas de “Tidy” para poder modificar las tablas de manera en la que los datos que teníamos agrupados en varias columnas pudieran pasar a ser variables en una sola columna llamada “categoría”. Estas funciones permitieron hacer este arreglo y aparte tener dos columnas más, una que definiera alguna característica específica de los valores de la columna de categoría y otra con la cantidad de observaciones que tenemos por cada categoría y cada grupo.

De igual manera usamos funciones que graficaran las tablas, estas funciones están creadas para separar por región cada departamento y clasificarlo. Luego de eso identifica cada característica y hace una gráfica distinta por cada característica y sus respectivos grupos.

**Exploración (EDA)**

Como un método preliminar de la examinación de la data se decidió crear una función que graficara las diferentes bases de datos, de esta manera podríamos observar datos a grosso modo, de las regiones del territorio nacional guatemalteco. Entonces, en base a este método se pudo observar los datos de una manera visual que nos ayudara a poder identificar datos o situaciones anómalas.

**Feature Engineering**

Dado que la información de las tablas las estamos analizando al unir los departamentos en sus respectivas regiones, se decidió normalizar la data para poder obtener los datos en las tablas de una manera en la que se pudieran analizar de buena manera. Ya que no todas las regiones cuentan con un número similar de departamentos, y, por consiguiente, de personas, la normalización ayudó a obtener datos que estuvieran estandarizados ya que de otro caso se pudo haber visto datos muy altos en determinadas regiones -como la suroccidental que son 6 departamentos- entonces el análisis estaba basado en cantidades irregulares de personas.

Esta normalización nos ayuda a tener cantidades de personas estandarizadas por región y poder llevar a cabo un análisis más exacto, la forma en la que realizó fue dividiendo el total de observaciones por el porcentaje de la población que representaba cada región.

**Gráficas**

Imagen que contiene instrumentos de escritura, papelería

Descripción generada automáticamente

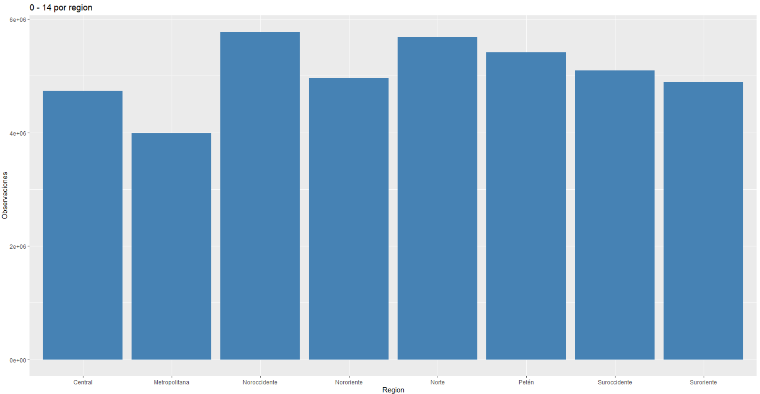
Mujeres de 15 años o más por número de hijos nacidos por región.

Imagen que contiene instrumentos de escritura, papelería

Descripción generada automáticamente

Imagen que contiene papelería, instrumentos de escritura

Descripción generada automáticamenteNivel Educativo de la Población de 4 años o más

Causa principal de Inasistencia de la Población de 4 a 29 años

Personas de 0 a 14 años por región

Imagen que contiene captura de pantalla

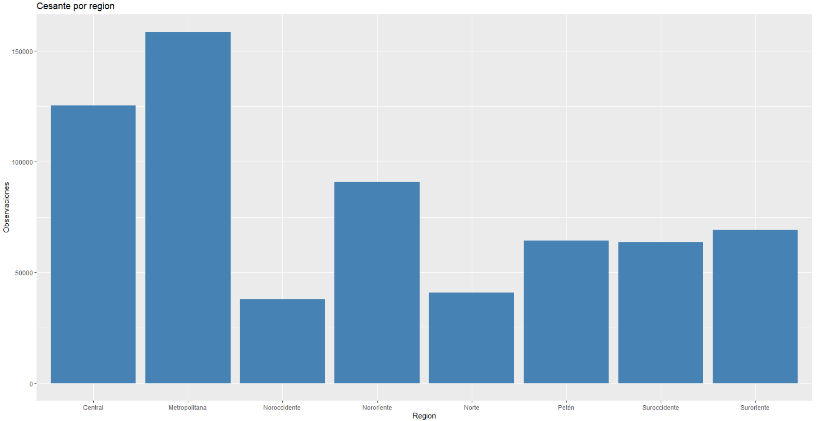
Descripción generada automáticamente

Personas de 30 a 64 años por región

Imagen que contiene captura de pantalla

Descripción generada automáticamente

Población aspirante.



Población Cesante

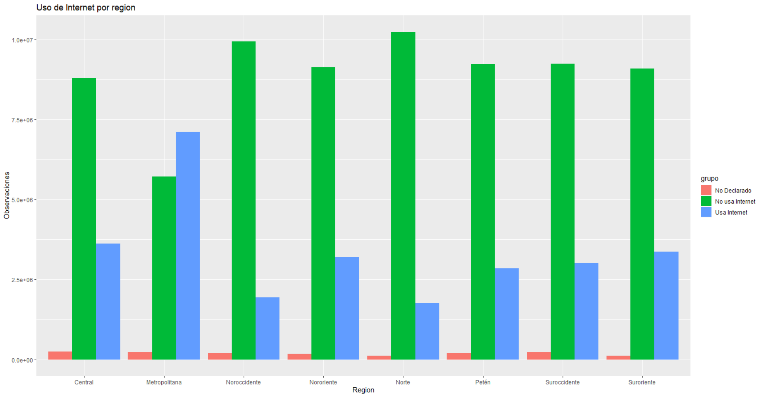
Imagen que contiene captura de pantalla

Descripción generada automáticamente

Población Económicamente Activa

Imagen que contiene captura de pantalla

Descripción generada automáticamente

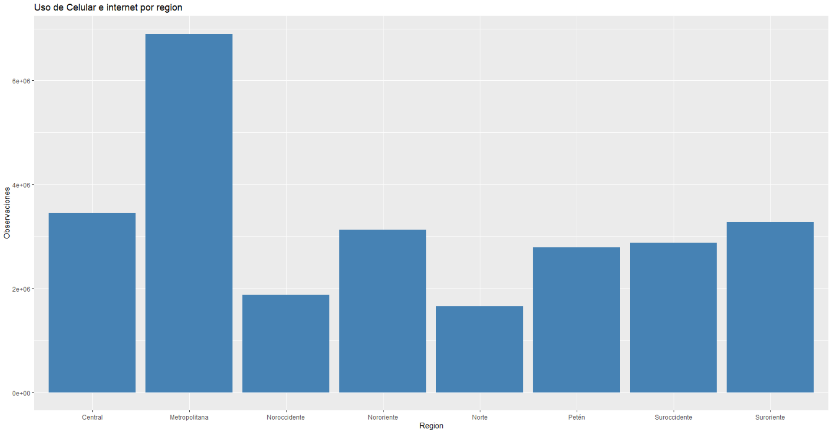
Población que usa Celular

Población que usa en Internet

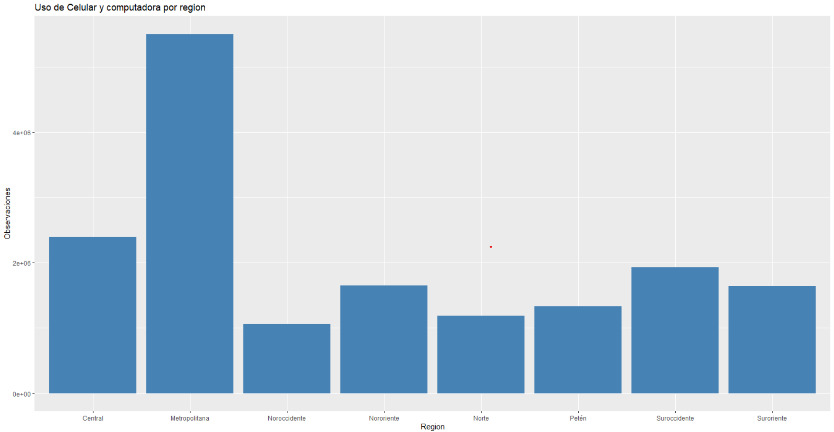
Imagen que contiene captura de pantalla

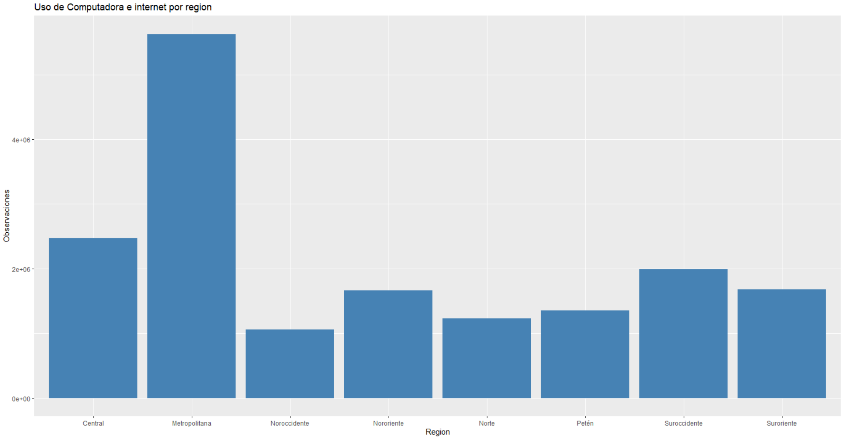
Descripción generada automáticamente

Población que usa Computadora

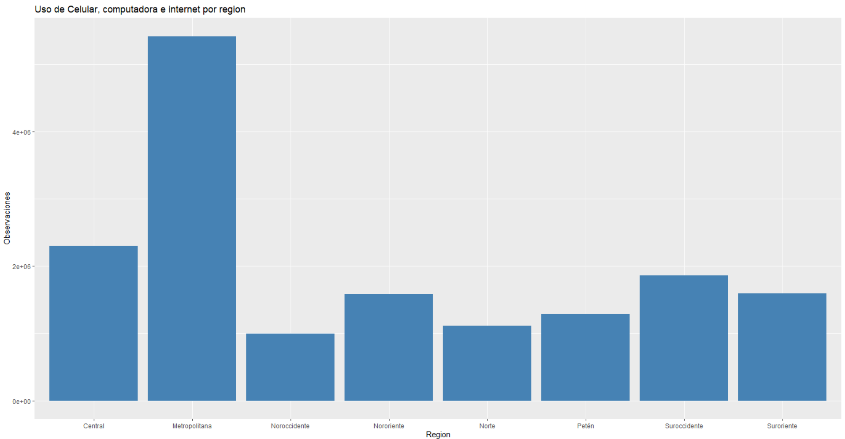


Población que usa Celular e Internet



Población que usa Celular y Computadora

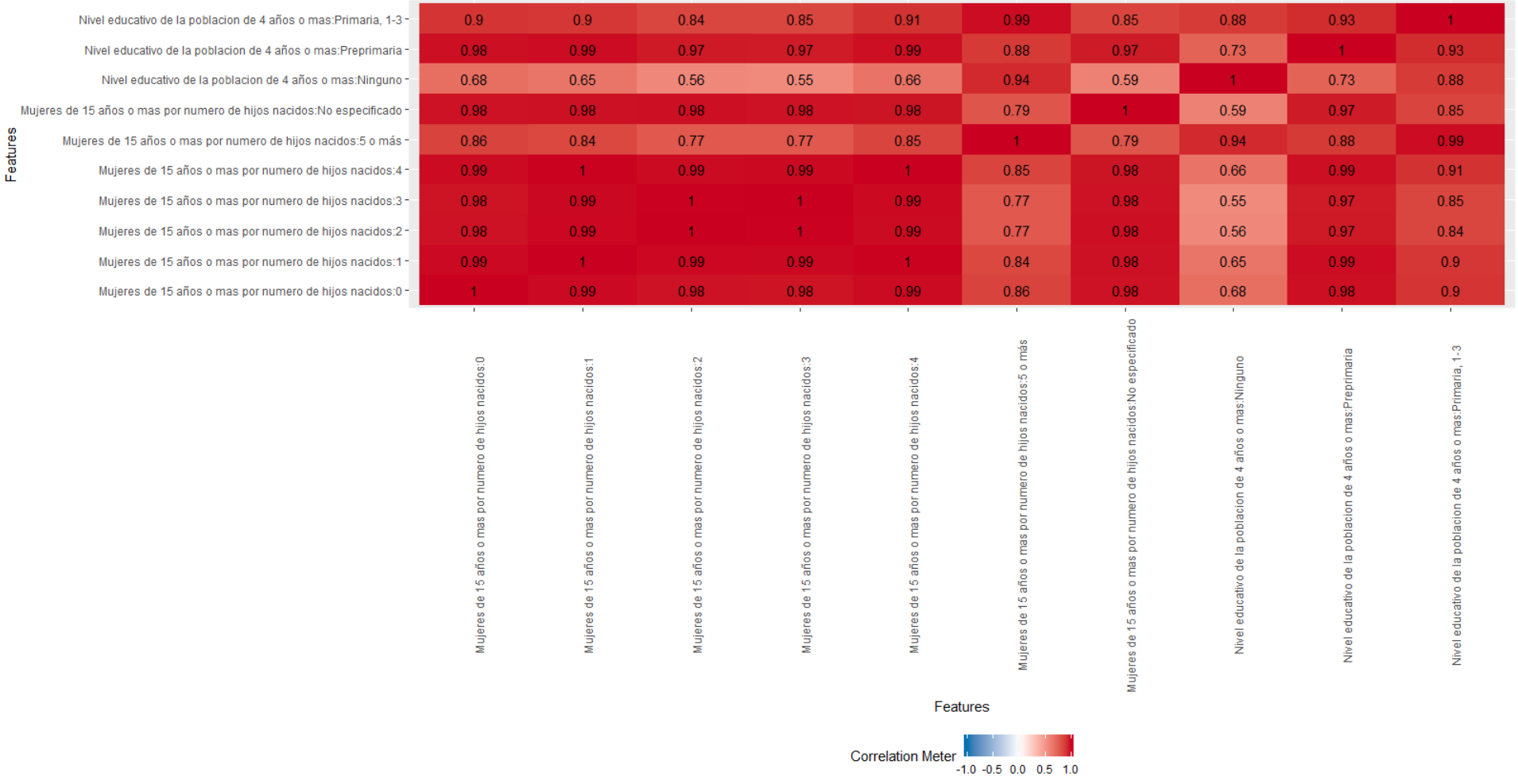
Población que usa Computadora e Internet



Población que usa Celular, Computadora e Internet

Imagen que contiene captura de pantalla

Descripción generada automáticamente**Correlación de Variables**

Correlación entre los niveles educativos y las causas de inasistencia a la escuela

Correlación entre el nivel educativo y la cantidad de hijos de mujeres de 15 años en adelante.

Imagen que contiene captura de pantalla

Descripción generada automáticamente

Correlación entre la inasistencia de las personas a la escuela con las poblaciones económicamente activas y las personas de 0 a 14 y 30 a 64.



Correlación entre el uso y falta de uso de dispositivos y herramientas tecnológicas y la edad de 30 a 64 años.

**Hallazgos**

Al haber realizado el análisis de la data y examinar cada región contra la metropolitana y el resto pudimos encontrarnos con resultados interesantes:

* El área metropolitana se caracteriza porque hay más mujeres mayores de 15 años con hijos que tienen, mayormente, de 0-3 hijos. En las regiones de noroccidente y norte podemos ver que la mayoría de las mujeres de 15 años o más con hijos tienen, mayormente, de 5 hijos en adelante.
* El área metropolitana tiene el mayor índice en el nivel educativo de niños mayores de 4 años desde los básicos para arriba (diversificado, universidad, posgrado) y las regiones norte y nororiente tienen niveles alarmantes de falta de educación. Hay un dato intrigante, la región metropolitana tiene la menor proporción de personas de niveles educativos de primaria de 1-3 y 4-5, en primaria 6 se convierte en promedio.
* Los motivos de inasistencia escolar en todas las regiones, a excepción de la metropolitana, son falta de dinero y otras causas, mientras que en la región metropolitana la razón principal de inasistencia escolar es necesidad de trabajar.
* La región metropolitana es la que menor cantidad de niños de 0-14 años tiene.
* La región metropolitana es la que mayor cantidad de adultos de 30-64 años.
* En todas las regiones la cantidad de personas que no pueden usar computadora es mayor que la cantidad de personas que sí pueden usar, pero podemos ver que el área que más puede usar computadora es el área metropolitana, el resto de las áreas tienen un gap mayor.
* Al examinar el uso de internet, se observa que el área metropolitana es el que más lo usa y el resto de las áreas no, en el área norte es donde hay una mayor diferencia entre no usar y usar internet. Lo mismo se puede observar con el uso del celular.
* Únicamente la región metropolitana es la que tiene más personas tiene que usa alguna combinación de internet, celular y computadora. En el resto de las regiones la cantidad de personas que no usa algunas de las combinaciones de internet, celular y computadora es menor a la cantidad de personas que sí los usa.
* Guatemala tiene el mayor nivel de personas económicamente activas y la región noroccidente es la que menos personas tiene.

**Conclusiones**

Al poder examinar a fondo los datos que se obtuvieron de el “XII censo de población y VII censo de vivienda” y poder analizar las 8 regiones del territorio guatemalteco pudimos encontrar lo siguiente:

* La región metropolitana es la que tiene un mayor nivel de desarrollo en educación y tecnología.
* La región metropolitana tiene una tasa de natalidad menor que el resto de las regiones, esto mismo lleva a que el nivel de niños menores de 0-14 sea menor que en el resto de las regiones.
* Las regiones norte y noroccidente son las que tienen los niveles más bajos de analfabetismo, educación, población económicamente activa y tecnología. Tienen los niveles más altos en natalidad por parte de madres de 15 años en adelante cuando son 5 hijos o más.
* En los indicadores que pueden considerarse “buenos” el área metropolitana tiene los mejores índices que el resto de las áreas, pero de igual manera los indicadores no son los más altos.

Luego de haber realizado y analizado las gráficas de correlación podemos asumir también algunas conclusiones:

* Existe una correlación fuerte entre la cantidad de hijos nacidos y el nivel de educación de preprimaria y primaria 1-3.
* Se descubrió que no hay una correlación entre la falta de escuelas e institutos con el nivel de educación
* La falta de dinero es la causa que más afecta al nivel educativo de primaria 1-3, 4-5 y 6 y poco a poco afecta menos al resto de niveles de educación. La causa de tener que trabajar es la que más afecta a la inasistencia en niveles superior de educación.
* Se puede encontrar una relación alta entre tener de 30 años en adelante y usar alguna de las herramientas de tecnología -internet, celular y computadora-.

**Recomendaciones y Siguientes pasos**

Recomendaciones para aquel que desee hacer un trabajo similar al nuestro: Si alguien deseara hacer algo como lo hemos hecho nosotros, tenemos ciertas recomendaciones que consideramos que podrían ser de mucha utilidad para poder hacer esto con mucha facilidad y no quitar grandes cantidades de tiempo de su horario.

* Tener bases de datos que se encuentren previamente “limpias” ya que de esta manera no tenemos que perder tanto tiempo usando herramientas para dejar las tablas en formato “Tidy”. Sabemos que esto no depende de nosotros porque la información nos la provee alguien más, pero de igual manera es de mucha ayuda y nos ahorra tiempo tener tablas ya ordenadas y mejor preparadas.
* Saber cómo crear funciones en R desde 0 es algo que consideramos extremadamente importante, ya que al saber cómo crear una función puedo hacer funciones que se adapten específicamente para los problemas e inconvenientes a los que me enfrento arreglando la base de datos. Entonces si conozco cómo hacer una función usando funciones que ya están en R o en paquetes específicos en poco tiempo puedo hacer yo una que solucione un problema de manera directa y no necesito buscar 3 o 4 funciones diferentes para llegar a un fin específico.
* Si sabemos específicamente qué buscamos nuestro trabajo puede ser muchísimo más fácil, al tener un objetivo principal en el horizonte podemos no gastar tiempo buscando cosas que realmente no son relevantes e ir directo al grano. Nosotros no tenemos un objetivo específico, entonces en nuestra búsqueda podemos encontrarnos con datos interesantes, pero de igual manera podemos encontrar cosas que no importen.

Se considera que los siguientes pasos a seguir luego de haber llevado a cabo el trabajo investigativo, se considera que con esta información las autoridades pertinentes podrían considerar los datos para actuar en base a ellos.

Con todos los resultados es posible observar el estado de diferentes regiones del país, por lo que se podría buscar arreglar factores que puedan ser influyentes en que el desarrollo de dichas regiones sea subóptimo.

**Anexos**

Diccionario de variables u otros términos:

* Código: La representación categórica de un departamento en las tablas de estadísticas departamentales.
* Departamento: El nombre del departamento guatemalteco de cual se realizó la observación.
* Característica: Grupo “Macro” o superior de una categoría de la cual se realizó una observación.
* Grupo: Columna que indica el grupo particular de la característica a analizar, en caso de que este sea “No hay grupo”, significa que únicamente es una característica principal del departamento u municipio.
* Observaciones: El número de personas que se observaron por censistas en el año 2018 para la respectiva característica y para el respectivo grupo en caso haya.
* Municipio: El código del municipio dentro del departamento que se indica en la columna departamento.
* Total de personas: El total de personas que se encontraron dentro del departamento en el 2018.
* Lugar de nacimiento (Lugar de residencia en 2013): El lugar de nacimiento (residencia en 2013) de los individuos entrevistados, estos pueden ser:
  + En el mismo municipio.
  + En otro municipio.
  + En otro país.
  + No especificado.
  + No había nacido \*(Lugar de residencia en 2013)
* Dificultades: Esta característica representa dificultades para ver, cuidado personal, recordar, oír, comunicar o caminar. Se desglosa en los siguientes grupos:
  + No declarado.
  + Con dificultad.
  + Sin dificultad.
* Total de mujeres en edad fértil: Número de mujeres que se encuentran en edad fértil en cada departamento.
* Mujeres de 15 años o mas por número de hijos nacidos (sobrevivientes): Esta característica representa el número de mujeres que tienen cierta cantidad de hijos nacidos (sobrevivientes) por cada departamento, los grupos para esta característica son:
  + 0
  + 1
  + 2
  + 3
  + 4
  + 5 o más
* Nivel educativo de la población de 4 años o más: Representa el último nivel educativo obtenido por el individuo censado, las cuales pueden ser:
  + Ninguno
  + Preprimaria
  + Primaria, 1 – 3
  + Primaria, 4 – 5
  + Primaria, 6
  + Básico
  + Diversificado
  + Licenciatura
  + Maestría y Doctorado
* Causa principal de Inasistencia escolar de la población de 4 a 29 años: Representa las razones por las que las personas de estas edades no estudian, las cuales son:
  + Falta de dinero
  + Tiene que trabajar
  + No hay escuela, instituto o universidad
  + Los padres/pareja no quieren
  + Quehaceres del hogar
  + No le gusta/No quiere ir
  + Ya terminó sus estudios
  + Otra causa
  + No especificado
* Población de 7 años o más: Población de individuos en el departamento que tienen 7 años o más que se distinguen en los siguientes grupos:
  + Analfabetas
  + Alfabetos
* Asistencia escolar 7 años o más: Población de individuos de 7 años o más que asisten al colegio que se distinguen en los siguientes grupos:
  + Asiste
  + No asiste
* Lugar de Estudio: Define el lugar en el que cada individuo estudia:
  + En el mismo municipio
  + En otro municipio
  + En otro país
  + No especificado
* Población de 7 años o más: Población de individuos en el departamento que tienen más de 7 años.
* Población Económicamente Activa: La población que trabaja y que genera ingresos:
* Población Ocupada: Población que tienen un empleo remunerado o bien se encuentran sin trabajar, pero tienen un vínculo con su empleo.
* Cesante: Personas que tienen experiencia laboral y buscan trabajo
* Aspirante: Personas que no tienen experiencia aboral y buscan conseguirla.
* Distribución de hogares por área urbana: Porcentaje de hogares que se encuentran ubicados en un área urbana
* Distribución de hogares por área rural: Porcentaje de hogares que se encuentran ubicados en un área rural
* Promedio de personas por hogar: Promedio de personas que viven en un determinado hogar
* Promedio de personas por dormitorio: Promedio de personas que duermen juntas en un dormitorio dentro de un hogar
* Hombre: Total de hombres del total de personas
* Mujer: Total de mujeres del total de personas
* 0-14: Total de personas que tienen edades entre 0 y 14 años
* 15-29: Total de personas que tienen edades entre 15 y 29 años
* 30-64: Total de personas que tienen edades entre 30 y 64 años
* 65-84: Total de personas que tienen edades entre 84 y 65 años
* 85: Total de personas que tienen edades mayores a los 85
* Grupo Quinquenal: Grupos que se agrupan de 5 en 5 años:
  + 0-4
  + 5-9
  + 10-14
  + 15-19
  + 20-24
  + 25-29
  + 30-34
  + 35-39
  + 40-44
  + 45-49
  + 50-54
  + 55-59
  + 60-64
  + 65-69
  + 70-74
  + 75-79
  + 80-84
  + 85-89
  + 90-94
  + 95-99
  + 100-más
* Urbano: Del área urbana
* Rural: Del área rural
* Parentesco con el jefe del Hogar: Relación una persona tiene con el jefe de hogar:
  + Jefe(a) de hogar
  + Esposa(o) o pareja
  + Hija(o) o hijastra(o)
  + Nuera o yerno
  + Nieta o nieto
  + Hermana o hermano
  + Madre o padre
  + Suegro o suegra
  + Cuñada o cuñado
  + Otra(o) pariente
  + No parientes
* Total, de personas en viviendas colectivas y situación de calle: Personas que actualmente viven en la calle o están viviendo en un hogar donde viven varias personas en situaciones similares.
* Población de 10 años o más: Población de personas que tienen 10 años o más
* Soltera(o): No está casada(o)
* Unida: Vive con alguien en una relación similar al matrimonio, pero sin estar oficialmente casados
* Casada(o): Casada o casado
* Separada(o): No está divorciada(o) pero no convive con la persona
* Divorciada(o): Legalmente divorciada(o) de la pareja
* Viuda(o): El esposo o esposa fallecieron
* Total de viviendas: El total de viviendas que se encuentran dentro del departamento/municipio.
* Total de viviendas particulares: El número total de viviendas que tienen acceso independiente las personas que viven dentro de ella, no siendo necesario pasar por otra vivienda para entrar a la del individuo.
* Total de viviendas colectivas: El número total de viviendas donde un conjunto de personas residen por motivos de salud, reclusión, religión u otro motivo.
* Tipo de vivienda particular: El tipo de vivienda particular que se observa, puede tomar los siguientes valores:
  + Casa formal
  + Apartamento
  + Cuarto de casa de vecindad
  + Rancho
  + Improvisada
  + Otro
  + Particular no registrada
* Tipo ocupación viviendas particulares: Menciona el nivel de ocupación de la vivienda particular que se observa, las cuales pueden ser:
  + Ocupada
  + De uso temporal
  + Desocupada
  + Rechazo total o parcial
* Material predominante pared: Este rubro indica el material predominante en la pared de la vivienda, los cuales pueden ser:
  + Ladrillo
  + Block
  + Concreto
  + Adobe
  + Madera
  + Lamina metálica
  + Bajareque
  + Lepa, palo o caña
  + Material de desecho
  + Otro
  + No especificado
* Material predominante techo: Este rubro indica el material predominante en el techo de la vivienda, los cuales pueden ser:
  + Concreto
  + Lamina metálica
  + Asbesto cemento
  + Teja
  + Paja, palma o similar
  + Material de desecho
  + Otro
  + No especificado
* Material predominante piso: Este rubro indica el material predominante en el piso de la vivienda, los cuales pueden ser:
  + Ladrillo cerámico
  + Ladrillo de cemento
  + Ladrillo de barro
  + Torta de cemento
  + Parqué o vinil
  + Madera
  + Tierra
  + Otro
* Uso de celular(internet|computadora): Este rubro indica el uso de celular(internet|computadora) por parte de los individuos censados, este uso se desglosa en tres grupos:
  + Usa celular(internet|computadora)
  + No usa celular(internet|computadora)
  + No Declarado
* Uso de celular, computadora e internet: Rubro que cuenta los individuos que utilizan el celular, la computadora y el internet en cada departamento/municipio.
* Uso de celular y computadora: Rubro que cuenta los individuos que utilizan el celular y computadora en cada departamento/municipio.
* Uso de celular e internet: Rubro que cuenta los individuos que utilizan el celular e internet en cada departamento/municipio.
* Uso de computadora e internet: Rubro que cuenta los individuos que utilizan la computadora e internet en cada departamento/municipio.
* Pueblo de pertenencia: El pueblo (ascendencia) al cual pertenecen los individuos censados.
* Comunidad lingüística: La comunidad lingüística a la que pertenecen los individuos censados.
* Idioma en el que aprendió a hablar: El primer idioma que el individuo utilizo para comunicarse con los demás.