**Nivel de Pobreza de los Hogares Costarricenses**

**Integrantes del grupo**

Macarena Flecha

Alain Kräupl

Leonardo Carabajal



**Introducción**

Muchos programas sociales tienen dificultades para brindarle la ayuda suficiente a las familias correctas, principalmente por la predominancia de trabajo en negro, ya que no permite tomar los ingresos salariales como principal medida de economía del hogar.

En América Latina, un método popular utiliza un algoritmo para verificar la calificación de ingresos. Se llama Proxy Means Test (o PMT). Con PMT, las agencias usan un modelo que considera los atributos observables del hogar de una familia como el material de sus paredes y techo, o los activos encontrados en el hogar para clasificarlos y predecir su nivel de necesidad.

Si bien esto es un avance en el ámbito, la precisión suele ser un problema a medida que crece la población. La principal falencia de este algoritmo es que se focaliza en clasificar a los no pobres por sobre los pobres.

Se propone crear nuevos métodos más allá de la econometría tradicional, basados ​​en un conjunto de datos relacionados con las características de los hogares costarricenses, los cuales podrían ayudar a mejorar el rendimiento de PMT. Para lograr esto el [**BID (la mayor fuente de financiamiento para el desarrollo de América Latina y el Caribe)**](https://www.iadb.org/en) nos suministró de un conjunto de datos de índice de pobreza de Costa Rica para que se intente mejorar el método tradicional usado hasta ahora.

Más allá de Costa Rica, muchos países enfrentan el mismo problema de evaluar incorrectamente las necesidades sociales. Creemos que con nuevos algoritmos se podría implementar en otros países del mundo.

## **Análisis Estadístico de Variables**

### **Selección de variables de interés**

Hicimos una calificación de las variables considerándolas como mayormente Causales o Consecuentes de pobreza, considerando que cuestiones relacionadas con los integrantes de las familias y sus cualidades como edad, dependencia, educación, edad, género podrían ser consideradas causas; y cuestiones relacionadas con descripciones físicas de la vivienda y tenencia de bienes materiales, consecuencias. No tenemos una opinión muy marcada sobre cómo calificar a las variables de composición del hogar y la relación familiar que los une por lo que decidimos obviar esas variables de la elección. A su vez por supuesto existe la clasificación de De Hogar o Individual, que indica si las variables caracterizan a la persona entrevistada en ese registro o al hogar en general.

Dicho esto, elegimos analizar dos variables consideradas Causa y dos Consecuencia: dependency (De Hogar), instlevel (Individual), overcrowding (De Hogar), TipoVivienda (De Hogar), además de v2a1 (De Hogar) por considerarla de sumo interés, la cual compararemos contra todas las demás.

Para el caso de las variables calificadas como De Hogar, como tienen valor único para todo el hogar, trabajamos con un único registro por cada IdHogar, ya que si considerásemos todos los registros influiría la cantidad de entrevistados (de registros del dataset) en los parámetros generales. En nuestro caso decidimos tomar por cada IdHogar el registro correspondiente al Jefe de hogar, es decir parentesco = 1, pero podría haberse tomado otro criterio.

De nuestro análisis podemos resumir lo siguiente para cada variable elegida:

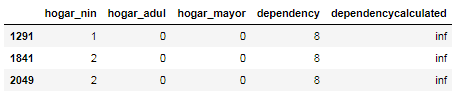
#### **Dependency**

Dependencia representa a la razón entre las personas consideradas dependientes por sobre las independientes. Se califica como dependiente a una persona solo por su edad, pero podrían considerarse otros factores como discapacidad o estudios en curso. Para el primer caso, si bien se tiene el dato de si hay discapacitados en la familia, no se sabe ni cuántos ni de qué edad, por lo que no se puede incorporar esto en la fórmula.

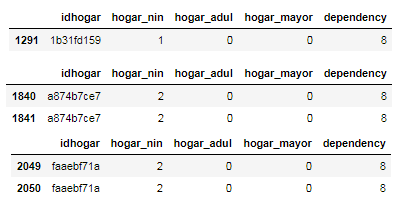
Si bien el concepto al que representa es del tipo numérico (continua) se ve que los datos son categóricos por tener valores del tipo "yes", "no". Como no se sabe qué significan estos datos pero sabiendo que esta variable se calcula a partir de otras, se efectúa el cálculo y compara contra el valor real. Como esta comparación fue muy positiva es que se decide descartar la variable que viene en el dataset y calcularla por cuenta propia.

En cuanto a los valores considerados típicos, estos toman valores entre 0 y 2,5 y se corresponden con el 97,74% del total de mediciones, el máximo valor por su parte llega a 8.

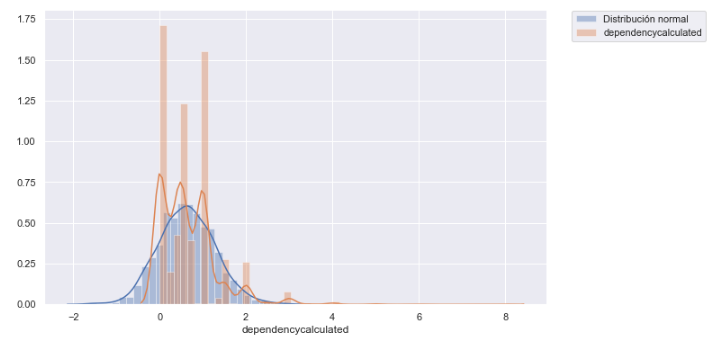
Existen a su vez 3 hogares cuyo cálculo no es factible de hacer por tener 0 adultos, incluso tienen 0 adultos mayores!! Si se busca el valor original que traía la variable, no es equivalente a la calculada, además de no concordar con su significado en cuanto a la composición familiar. Si bien se decice excluir, simplemente por la imposibilidad de dividir por 0, y por ser solo 3 hogares, claro está que estos casos deberían tratarse con cierta particularidad, considerando la importancia de darle ayuda económica a un hogar con estas características si es que efectivamente las presenta, las consecuencias de clasificar mal un hogar de este tipo podrían ser muy perjudiciales.



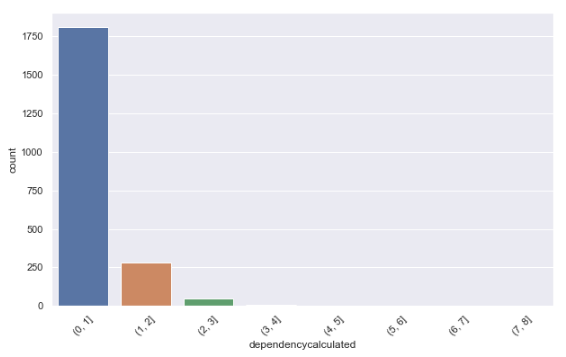
##### ¿Y si buscamos qué datos tienen estas mismas variables en las demás entrevistas de los mismos hogares?



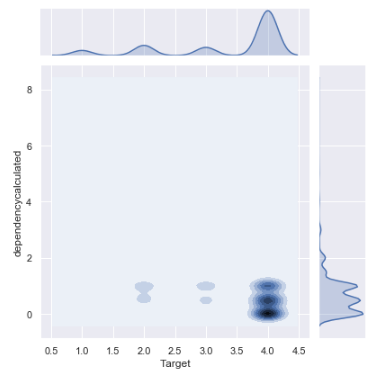
Efectivamente no son datos consistentes, no hay adultos mayores pero la dependency indica relación de 8 entre estas columnas, es correcto excluir estos casos.



En cuanto a las distribuciones que podrían representar esta variable, se puede dar por sentado que no tiene similitud alguna con una normal. La distribución en este caso presenta picos o altibajos, y es totalmente asimétrica.



Predominan las dependencias con valores entre 0 y 1, y como era de esperarse los valores atípicos ni llegan a verse en los gráficos para la escala planteada.



Al evaluar cómo se comporta el **Target** para los valores de la dependencia, se llega a la conclusión de que la mayoría de la distribución implica dependencias bajas y Target = 4. Ya se sabe que en sí la mayoría de las dependencias implicaban valores bajos, lo que es novedad es a qué clase pertenecen.

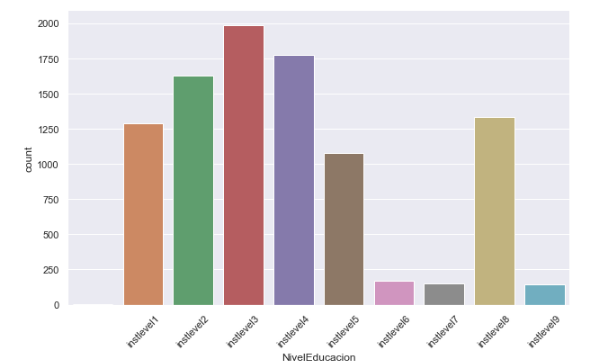
Se concluye que a medida que el Target mejora (aumenta) la dependencia disminuye considerablemente, lo cual podría ser lógico desde el negocio en pensar que las personas dependientes no tienen ingresos, solo generan gastos, por lo que se espera que sean más pobres. Recordar que la variable dependencia no indica cantidad sino razón, es decir que un valor alto no necesariamente implica muchos niños/ancianos sino es la comparación contra la cantidad de adultos.

#### En cuanto a la relación con el tamaño del hogar, **tamhogar**, no existe una relación aparente entre las variables, aunque se puede apreciar inicialmente una tendencia a suba. Se puede apreciar que los promedios son irregulares a medida que el tamaño aumenta. Ahora, si se descartan los valores atípicos para los valores del tamaño de hogar los cuales van de 7 en adelante según el criterio de 1.5 \* IQR ya si podría dar por sentado una dependencia entre estas variables.

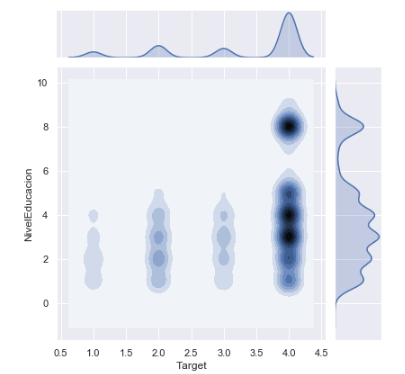
#### **Instlevel**

En si, esta no es una variable sino un conjunto de variables del tipo bandera. Para facilitar el trabajo se decide unificarlas.

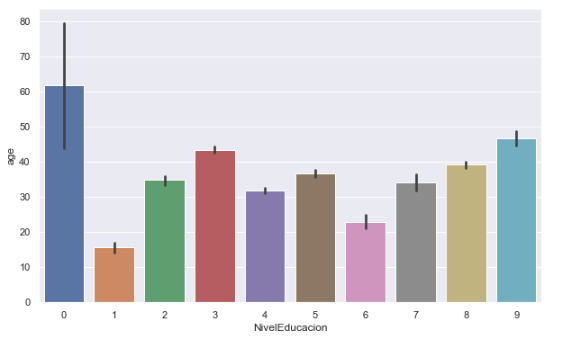
Por ser categórica, este análisis difiere un poco de la variable anterior, ya que, por ejemplo, no es factible comparar la distribución contra una normal o considerar valores atípicos.



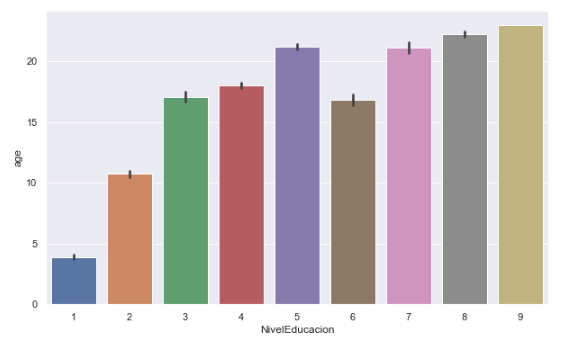
Si por supuesto se considera la cantidad de veces que los valores ya predefinidos se repiten. Tener presente que esta sí describe al individuo, por lo que se trabaja con el dataset completo. Se detectan que 3 valores en un total de 9557 están vacíos lo que es muy poca cantidad, estos registros podrían eliminarse. Como ya existe un valor que significa NO educación no podría deducirse este significado para un vacío. A su vez como son valores individuales, no es posible ir a buscarlos a otros registros del mismo hogar.



Al evaluar la distribución de esta variable respecto a **Target**, se observa que quienes pertenecen a una mejor clase son aquellos con mayor nivel de estudios (por supuesto que hemos reemplazado instlevel1 por 1, instlevel2 por 2 y así sucesivamente), mientras que para las demás clases rondan los mismos niveles, no así las densidades. Además es muy notoria la poca distribución para los niveles 6 y 7, lo cual podría ser lógico a nivel de negocio, podría ser un país en donde no son comunes las tecnicaturas o los secundarios técnicos.



#### En cuanto a las edades, se esperaría una distribución creciente, ya que los niveles educativos por supuesto están directamente relacionados con la edad, sin embargo esto no sucede.



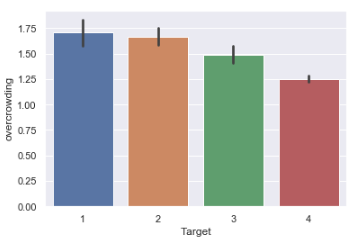
¿Ahora, si se considera solamente las edades implicadas con la educación efectuada en tiempo y forma y descartamos a los adultos y adultos mayores? Claramente los menores y adolescentes si siguen una educación esperada (no olvidar que los valores 6 y 7 por ser tecnicaturas se espera que se salgan del modelo), lo que podría deducirse de que por ejemplo el gobierno ha estado trabajando considerablemente fomentando la a educación en los últimos años.

#### 

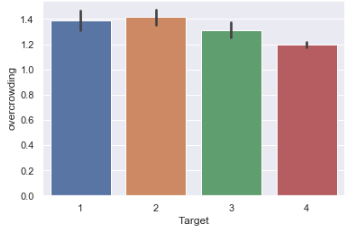
#### **Overcrowding**

Esta variable es una medida de sobrepoblación en un espacio dado, el hogar en este caso. Se espera que esté fuertemente ligada a la pobreza ya que esta condición no suele ser una elección de vida. Afecta considerablemente a la calidad de vida, desde cuestiones de falta de privacidad hasta incluso higiene, lo cual conlleva a problemáticas de salud. El cálculo depende de quien lo evalúe, algunos consideran la cantidad de habitaciones y otros el metro cuadrado por la cantidad de personas. En este caso se hizo con la cantidad de dormitorios sobre el total de personas, y si bien al efectuar el cálculo la gran mayoría de los datos coinciden con el valor que ya viene en el dataset, algunos pocos difieren. Se decide esta vez tomar el valor tal cual viene ya que a diferencia de dependency donde la fórmula estaba en la descripción de la variable, para overcrowding no lo está.

Al comparar la distribución contra **Target** se puede deducir que ha tenido cierto peso al momento de decidir en qué clase caía el hogar ya que a medida que la clase implica una mejoría económica esta medida va disminuyendo.



Al comparar la distribución contra **Target** se puede deducir que ha tenido cierto peso al momento de decidir en qué clase caía el hogar ya que a medida que la clase implica una mejoría económica esta medida va disminuyendo.



Es importante recordar que este gráfico representa un promedio de los valores por lo que es necesario siempre rehacerlo quitando los outliers o los valores que uno como analista considere demasiado grandes. En este caso la distribución se suaviza notoriamente, ya no tiene tanto peso en la decisión final esta medición.

#### **TipoVivienda**

Los tipos de vivienda posibles son:

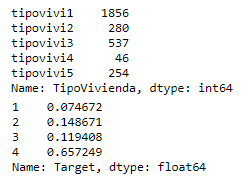
1: casa propia y totalmente pagada

2: propio, pagando a plazos

3: rentada

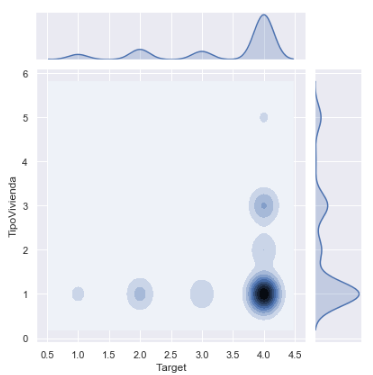
4: precaria

5: otro (asignado, prestado)



Teniendo en cuenta que el objetivo de este estudio es clasificar a los hogares por su grado de pobreza, hay un fuerte interés por el tipo de vivienda 4 y deducir qué otros factores condicionan a un hogar a considerarlo precario.

Claramente no es la variable de mayor peso porque por ejemplo un 22% de hogares son pobres, pero solo un 1% son precarios.



Sin embargo se cumple la lógica esperada de suponer que quienes tienen casa propia pertenecen a un mejor status económico, es decir clase 4.

#### 

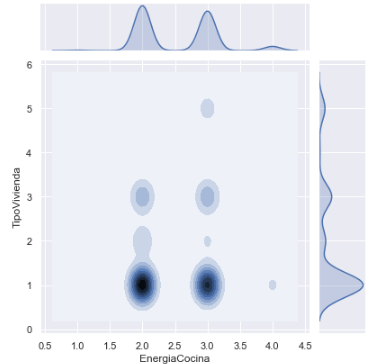
#### 

#### 

#### 

Los criterios por los que el encuestador califica a la vivienda entre estas categorías posibles es desconocido, se supone que los encuestadores han tenido una capacitación para ello pero podría existir cierto sesgo por tratarse de un valor con tendencia subjetiva.

se hizo una primer comparación contra las variables relacionadas a instalaciones y servicios, como son electricidad, agua, luz. Algunos análisis carecen de sentido por existir un alto porcentaje de hogares con un mismo valor, por ejemplo, la gran mayoría de los hogares presentan instalaciones de agua dentro, o electricidad pública.



En cuanto a la energía de cocina predominan dos valores, gas y electricidad prácticamente por igual, y ambos están fuertemente ligados a los tipos de vivienda 1, aunque hay una leve diferencia a favor de la electricidad.

#### 

#### 

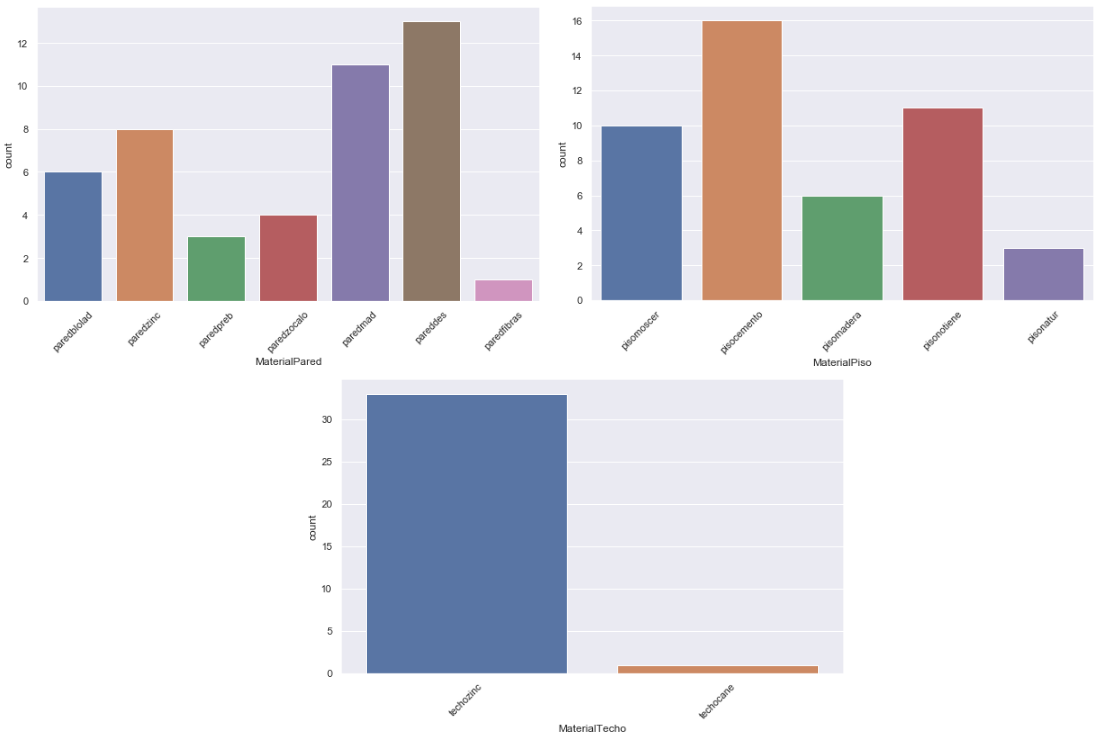
#### 

#### 

#### 

#### 

Según <https://www.larepublica.net/noticia/costa_rica_es_el_quinto_mas_expuesto_a_desastres_naturales>, Costa Rica presenta riesgos de catástrofes sísmicas por lo que, más que las instalaciones, es importante considerar la construcción del hogar. Si bien identificar qué materiales son los esperados para hogares antisísmicos excede a este estudio, es importante mínimamente observar de qué material están hechos los hogares, focalizando en los considerados precarios, por lo que los gráficos presentados a continuación tienen el filtro de tipo de vivienda = 4.



**V2a1**

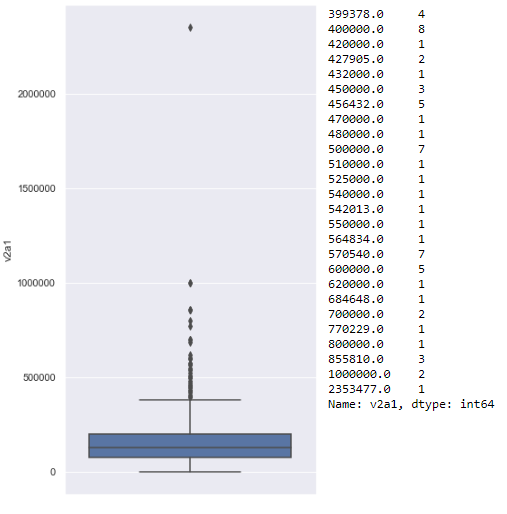
Esta variable representa a la renta del hogar. Considerando que la relación entre un hogar y una vivienda podría muchos a uno, no está clara la definición de la renta ya que todo el dataset está orientado al hogar, sin embargo la renta es una medida que describe a una vivienda. Como no hay información suficiente por no poder identificar viviendas entre los datos se dará por asumido una relación uno a uno.

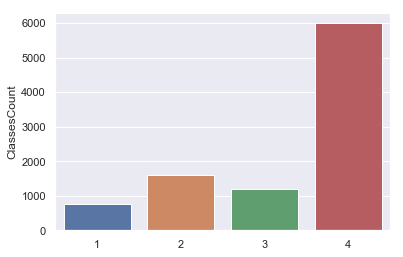
Interiorizándose en la calidad de los datos, existen No existen viviendas del tipo 3 (rentadas) con valores nulos o vacíos, lo cual es correcto. Sin embargo SI existen 274 hogares que siendo del tipo 2 declaran pagar renta. Como son demasiados y no se tiene acceso a más información sobre la entrevista podría pensarse que tal vez el tipo 2 signifique que la familia se está pagando su propia casa pero que actualmente está viviendo en una casa rentada. Son demasiados datos como para excluirlos por considerarlos un error.

Outliers

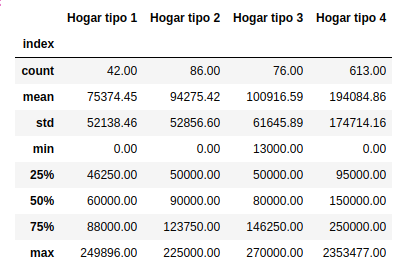
El monto medio de la renta mensual está centrado en torno a los ₡168.000(Colones). Además, existe una gran cantidad de hogares con renta nula, esto puede deberse al tipo de vivienda, en las que el hogar sea de propiedad de los integrantes.

Existen algunos pocos hogares con alquileres muy altos y alejados de la renta media por tipo de vivienda, estos podrían considerarse outliers, es decir, valores atípicos que podrían considerarse excluidos de los análisis ya que afectan el comportamiento general.



General vs Clasificado

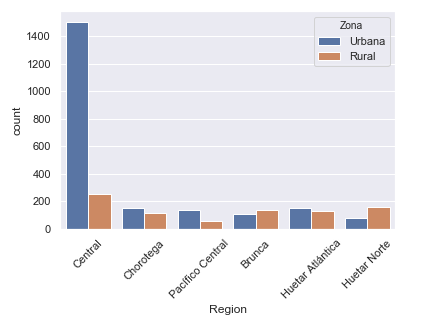
Si el análisis se realiza por nivel de pobreza comenzamos a notar grandes diferencias. La media general obtenida se ve muy afectada por los hogares de mejor clasificación, además de ser la mayoría, son los de mayor media mensual. En estos hogares, el valor promedio de renta es prácticamente el triple que el de su siguiente clase, haciendo que la media general sea muy superior a la que se observa por clase.

Como era de esperarse, la renta media se reduce a medida que la clasificación de los mismos es de mayor pobreza. 

Si bien la renta media de los hogares tipo 4 es superior a los demás, la desviación que se observa en esta clase es tres veces superior a las otras, algo que, luego de haber observado su gráfico de distribución, nos dice que los outliers terminan influyendo en ese incremento.

Otros aspectos a resaltar son que los hogares tipo 3 no poseen rentas en 0 y que la renta máxima del hogar tipo 1 es superior a los hogares tipo 2.

Análisis por región



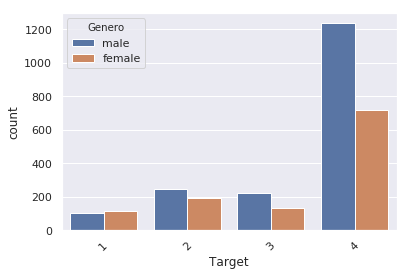
Al sectorizar el análisis por zonas geográfica y tipo de población son otras las conclusiones a las que podemos llegar.

Notamos la gran cantidad de hogares que habitan en zonas urbanas de la región central. El resto de los hogares se distribuyen de manera prácticamente uniforme a lo largo del resto del país.

Aspectos a resaltar sobre esta distribución es que la región Pacífico Central tiene más del doble de hogares en su zona Urbana y lo contrario pasa en Huetar Norte.

Cabe aclarar que, si bien la cantidad de hogares es un indicador del tamaño de la población, este último estará directamente relacionado con la cantidad de personas que habite cada hogar, aunque puede servir como aproximación de gran verosimilitud.

Volviendo al análisis de la renta, aunque ahora segmentado por región volvemos a justificar que el elevado promedio de la renta general se da por los montos que abona la gran mayoría de los hogares en la zona central.

Género del jefe del hogar

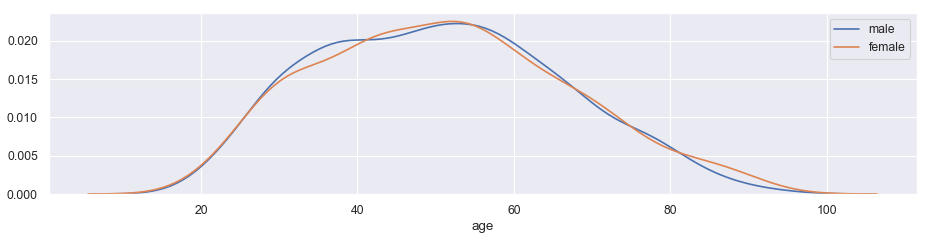
El censo económico realizado en el país arrojó que existen 3973 hogares, de los cuales 1163 son llevados adelante por mujeres.

Llevándolo al plano del análisis por clasificación de pobreza advertimos que los hogares con menores recursos, o al menos de menor renta, son llevados adelante por mujeres en su mayoría.

La brecha entre cantidad de hogares con jefes masculinos vs femeninos se ve reducida conforme se reduce el nivel de pobreza del mismo, a tal punto que en el máximo nivel, existen más hogares llevados adelante por mujeres (118 de género femenino vs 104 de género masculino).

Edad del jefe de hogar

Otro aspecto a destacar cuando analizamos los géneros de los jefes de hogar es el de la distribución de edades de los mismos. Los jefes de hogar masculinos comienzan a ejercer su rol a temprana edad, existe una mayor proporción de estos entre aproximadamente 28 y 40 años, a partir de ahí, salvo algún altibajo, la edad de los jefes de hogar se distribuye relativamente uniformes respecto del género hasta pasados los 80 años. Este último aspecto podría deberse, por ejemplo a la diferencia de esperanza de vida que existe entre los géneros.

******

### 

### **Conclusión**

El objetivo de esta primer etapa era descubrir y analizar el dataset que nos brindaron para poder entender cómo funciona el modelo estadístico que se estaba utilizando para las mediciones de niveles de pobreza que presentan las regiones de Costa Rica, este modelo es requerido para detectar Hogares con un nivel de pobreza alto para luego poder seleccionarlos y brindarles un apoyo económico acorde a cada situación.

El modelo actual tal como se menciona al comienzo del informe evalúa diferentes variables y en base a ello genera una métrica que indica el nivel de pobreza del hogar Costarricense.

Nosotros en nuestro análisis de las variables en una primera instancia detectamos que la mayoría eran condicionantes del nivel de pobreza que presentaban los Hogares, pero luego de profundizar un poco más observamos que las variables tendían a no condicionar el nivel de pobreza.

Por ejemplo la cantidad de personas por habitación no es una variable que podamos tomar en cuenta en este caso para determinar el nivel de pobreza ya que la mayoría de Hogares no supera 2 personas por habitación. Tal vez si el promedio de personas por habitación hubiese sido más alto en las encuestas podríamos tomarla como un condicionante directo del nivel de pobreza ya que esto implicaría que al no tener los recursos necesarios de acuerdo a la cantidad de habitantes de un hogar hace que exista sobrepoblación en habitaciones, quitando de algún modo privacidad, comodidad entre otros aspectos.

No es así con el nivel de educación, ya que si en este caso es un condicionante directo del nivel de pobreza en los hogares costarricenses ya que notamos que mientras menor nivel de pobreza de un hogar mayor es el nivel de educación al que pueden acceder los integrantes de los mismos. Podríamos decir que un integrante de un hogar que tiene un mayor nivel de educación posee más probabilidades de poseer un puesto laboral frente a otro con menor nivel de educación.

También la probabilidad de que un hogar que presente una jefa de hogar es condicionante a que sea un hogar con un nivel de pobreza mayor, ya que bien se sabe que las ofertas laborales para el sexo femenino siempre es menor a las ofertas del sexo opuesto.

En conclusión podríamos recomendar realizar análisis más profundos y contextualizados con problemas reales de la región donde se realizan las encuestas, ya que la pobreza es diferente en cada región, es decir que para cada región las variables tienen más o menos peso que otra ya que esto podría ser un gran condicionante al momento de indicar que un hogar presenta un nivel de pobreza bajo mientras el mismo debería haberse calificado con un nivel de pobreza más alto lo cual presenta un riesgo de mal calificación ya que no es lo mismo calificar a un hogar con recursos con un nivel alto de pobreza que calificar a un hogar como un hogar con nivel de pobreza alto como si tuviese un nivel de pobreza bajo ya que esto implicaría no brindar al hogar que realmente lo necesita.