PROYECTO PREPARACIÓN DE DATOS

ANÁLISIS NIVELES DE INGRESO Y COBERTURA EDUCACION POR DEPARTAMENTO AÑO 2019

Luis Omar García Amador

César Andrés Gaviria Cuevas

Laura Juliana Mora Páez

Laura Katherine Murcia Falla

Fabian Alexis Pallares Jaimes

Presentado para la clase de: GESTIÓN DE DATOS

Ingeniera Alexandra Pomares

FACULTAD DE INGENIERIA

DEPARTAMENTO DE INGENIERIA DE SISTEMAS

PONTIFICIA UNIVERSIDAD JAVERIANA

BOGOTÁ

2021

INTRODUCCIÓN

Nuestro análisis pretende indagar sobre la cobertura de la educación en el país, en busca de comprender la relación que existe entre el nivel educativo en cada departamento con los índices de ingresos económicos que enmarcan al país. De manera que se pueda determinar cómo influyen factores claves como la cobertura en educación y/o el nivel de matriculación escolar, en el desarrollo económico de los hogares.

Para el análisis, nos basamos en dos conjuntos de datos obtenidos de los portales de datos abiertos del gobierno y el archivo nacional de datos del DANE respectivamente. El primero corresponde a las estadísticas de educación por departamento del Ministerio de Educación Nacional (MEN) en el año 2019. Por otro lado, se cuenta con los datos de medición de pobreza monetaria y desigualdad del DANE, para el mismo año del estudio. Se obtiene entonces información relacionada con la cobertura por departamento de los distintos niveles de educación, la aprobación y reprobación, así como la tasa de deserción. De igual forma, el estudio la medición de pobreza monetaria y desigualdad, nos permite conocer datos relevantes sobre los hogares, con factores clave como el hecho de habitar en vivienda propia, arriendo o cualquier otra condición, así como las personas que componen el hogar, y el ingreso económico total de cada hogar

Este tipo de análisis puede aportar una visión clara de la necesidad de políticas educativas efectivas, que permitan fomentar la inclusión de niños y jóvenes en etapa escolar con el fin de mejorar a futuro sus condiciones de vida y disminuir la brecha de desigualdad que afecta al país.

Etapa 1: Carga y Exploración.

1.1. Entendimiento de los datos

Inicialmente se tenían dos sets de datos con 63 atributos, 25 correspondientes a los datos de medición de pobreza monetaria y desigualdad del DANE, y 38 a las estadísticas por educación primaria, básica y media por departamento del Ministerio de Educación Nacional (MEN). Se realizo un filtro de columnas con información no relevante para el estudio en curso (por ejemplo, sedes conectadas a internet, directorio de hogar, dominio, código fex de departamento, entre otras), de manera que los atributos quedaron reducidos a 20 en total, con los datos propios que se enfocan en el análisis planteado.

Adicional a ello, se realiza un filtro de filas en las estadísticas de educación con el fin de enfocar el estudio únicamente en el año 2019. Y finalmente se transforman atributos strings en double, tales como tasa de matriculación, Cobertura, Deserción, Aprobación, Reprobación, Repitencia, y el ingreso total por hogar, para ser tratado como factor numérico.

Con el fin de facilitar el entendimiento de los atributos, se realiza el renombre de las siguientes variables:

- p5000: NumCuartos
- p5090: TipoVivienda
- p5100: Amortización
- p5130: ArriendoHipotetico
- p5140: ArriendoReal

Finalmente se realiza la unión de ambos sets de datos, obteniendo los atributos necesarios.

1.2. Extracción y descripción inicial de los datos

Atributo	Procedencia	Tipo Dato Almacenamiento	Tipo Dato Conceptual	Descripción
AÑO	MEN	Integer	Numérico	Vigencia del indicador
CÓDIGO_DEPARTAMENTO	MEN	Integer	Cuantitativo Discreto	Código DANE del Departamento
DEPARTAMENTO	MEN	String	Cualitativo Nominal	Nombre del Departamento
POBLACIÓN_5_16	MEN	Integer	Cuantitativo Continuo	Población en edad teórica* de estudiar según proyecciones de población del DANE
TASA_MATRICULACIÓN_5_16	MEN	Double	Cuantitativo Continuo	Proporción de la población entre 5 y 16 años que se encuentra asistiendo al sistema educativo. Cuando las proyecciones de población del DANE no capturan adecuadamente los flujos migratorios internos, puede alcanzar valores mayores al 100%.

COBERTURA_NETA	MEN	Double	Cuantitativo Continuo	Es la relación entre el número de estudiantes matriculados en transición, primaria, secundaria y media que tienen la edad teórica* y el total de la población correspondiente a esa misma edad. Cuando las proyecciones de población del DANE no capturan adecuadamente los flujos migratorios internos, puede alcanzar valores mayores al 100%.
COBERTURA_BRUTA	MEN	Double	Cuantitativo Continuo	Es la relación entre el número de estudiantes matriculados en transición, primaria, secundaria y media respecto a la población en edad teórica* para cursar estos niveles. En algunos casos la demanda social es mayor a la población en edad teórica* para cursar educación preescolar, básica y media, explicada por estudiantes en extraedad, por lo que el indicador toma valores superiores al 100%.
DESERCIÓN	MEN	Double	Cuantitativo Continuo	Tasa de deserción intra - anual del sector oficial. Identifica la proporción de alumnos matriculados que por factores culturales, coyunturales o de prestación del servicio educativo, abandonan sus estudios durante el año lectivo.
APROBACIÓN	MEN	Double	Cuantitativo Continuo	Tasa de aprobación de estudiantes del sector oficial. Identifica el porcentaje de alumnos en educación preescolar, básica y media que aprueba de acuerdo con los planes y programas de estudio vigentes.
REPROBACIÓN	MEN	Double	Cuantitativo Continuo	Tasa de reprobación de estudiantes del sector oficial. Identifica el porcentaje de alumnos en educación preescolar, básica y media que reprueba de acuerdo con los planes y programas de estudio vigentes.
REPITENCIA	MEN	Double	Cuantitativo Continuo	Tasa de repitencia del sector oficial. Corresponde al porcentaje de alumnos matriculados en un año escolar en transición, primaria, secundaria y media que se encuentran repitiendo el mismo grado cursado el año anterior.
NumCuartos (P5000)	DANE	Integer	Cuantitativo Continuo	Incluyendo sale-comedor, ¿cuántos cuartos en total dispone este hogar?
TipoVivienda (P5090)	DANE	Integer	Cualitativo Nominal	L a vivienda ocupada por este hogar es: a. Propia, totalmente pagada b. Propia, la están pagando c. En arriendo o subarriendo d. En usufructo e. Posesión sin título (ocupante f. Otra)
Amortización (P5100)	DANE	Integer	Cuantitativo Continuo	¿Cuántos pagan mensualmente por cuota de amortización?

ArriendoHipotetico (P513)	DANE	Integer	Cuantitativo Continuo	Si tuviera que pagar arriendo por esta vivienda, ¿cuánto estima que tendrá que pagar mensualmente?
ArriendoReal (P5140)	DANE	Integer	Cuantitativo Continuo	¿Cuánto pagan mensualmente por arriendo?
NPER	DANE	Integer	Cuantitativo Continuo	Número de personas en el hogar
INGTOTUGARR	DANE	Double	Cuantitativo Continuo	Ingreso total de la unidad de gasto con imputación de arriendo a propietarios y usufructuarios.
POBRE	DANE	Integer	Cualitativo Nominal	Pobre 0 No pobre 1 Pobre
INDIGENTE	DANE	Integer	Cualitativo Nominal	Indigente 0 No indigente 1 Indigente

1.3. Análisis exploratorio de datos / Análisis de calidad

Después de la selección de atributos mencionada anteriormente, se procedió a juntar las tablas utilizadas con un nodo *joiner* dentro de KNIME mediante los atributos correspondientes al código del departamento.

Una vez juntadas las tablas en un solo conjunto de datos, se realizó un análisis exploratorio de datos, en donde se analizaron algunos aspectos como las medidas de tendencia central, métricas de dispersión, cantidad de valores nulos y valores atípicos para facilitar el entendimiento de estos datos. Algunas de estas medidas se exponen en la tabla presentada a continuación y en las visualizaciones posteriores a esta.

Nombre atributo	Mínimo	1er cuartil	Media	Mediana	3er cuartil	Máximo	Std. Dev.	Skewness	Kurtosis	N° nulos
POBLACIÓN_5_16	83130	218194	404400,66	299016	461988	1181531	316896,32678	1,484	0,98613	0
TASA_MATRICULACIÓN_5_16	0.8168	0,9052	0,926	0,94	0,967749	0,994446	0.054	-0,744	-0.622	0
COBERTURA_NETA	0,814564	0,9049	0,924	0,9383	0,963015	0,993342	0.053	-0,748	-0.603	0
COBERTURA_BRUTA	0,926231	1020565	918545,3264	1064933	1089631	1148539	374635.44	-2,016	2.145	0
DESERCIÓN	0,012632	0.0278	0,033	0,0325	0.03972	0,053201	0.01	-0,171	-0.413	0
APROBACIÓN	0,858552	0.88597	0,907	0,905	0.929093	0,975008	0.027	0,71	0.056	0
REPROBACIÓN	0,009037	0.05035	0,06	0,0645	0.075932	0,096673	0.022	-0,658	-0.278	0
REPITENCIA	0,007452	0.01164	0,02	0,0153	0.031925	0,036743	0.01	0,44	-1,442	0
p5000	1	3	3,3744	3	4	18	1,1974	0,3787	1,7307	0
p5100	98	320000	1186824,2024	560000	1000000	930000000	12923946,9596	53,6054	3507,6865	223860
p5130	98	200000	583207,7127	400000	600000	900000000	6091847,4498	90,3497	10317,6363	91865
p5140	9	290000	464740,8608	400000	520000	900000000	3297728,7054	235,867	61571,0133	139966
nper	1	2	3,2613	3	4	28	1,7888	1,3458	4,5229	0
ingtotugarr	0	900000	2343775,5191	1596333,3333	2827000	158000000	2712713,3747	6,1293	109,2911	0
pobre	0	0	0,2819	0	1	1	0,4499	0,9696	-1,0599	0
indigente	0	0	0,0652	0	0	1	0,2469	3,5217	10,4028	0

Al realizar un análisis de los datos obtenidos, revisando el valor mínimo, la media, mediana, el máximo y los valores de los cuartiles, podemos notar, a simple vista, una diferencia bastante amplia a nivel de la población entre los 5 y 16 años que acceden a educación, ya que esta varía entre las 83.130 y las 1.181.531 personas; por otro lado se puede observar que la mayoría de poblaciones con este acceso, oscilan entre las 218.194 y las 461.988 y se cuenta con un promedio de 404.400,66 y una mediana de 299.016 así como se cuenta con una desviación estándar de 31.896,32678. Con estos datos en mente, podemos llegar a considerar que el máximo y mínimo son valores atípicos, con la hipótesis que pueden pertenecer a poblaciones muy pequeñas o muy grandes respectivamente, o que existió algún error a la hora de ingresar los datos.

Pasando a mirar la tasa de matriculación que se tiene en estos intervalos de edad, tenemos que los datos se encuentran entre el 0,8168 y el 0,9945 de estudiantes matriculados, el primer cuartil se ubica en 0,9052 y el tercero en 0,967749, se presenta una media de 0,926, una mediana de 0,94 y una desviación estándar de 0,054. Con estos datos se puede destacar que la cantidad de jóvenes matriculados, en los diferentes departamentos del país, es alta ya que este porcentaje esta entre 81,68% y el 99,45%.

Pasando a la cobertura neta, se tiene que los datos oscilan entre 0,814564 y 0,993342, donde el primer cuartil se ubica en 0,9052 y el tercero en 0,967749, se tiene una media en 0,924, una mediana en 0,9383 y una desviación estándar de 0,053. Con esto podemos observar que realmente la cobertura es bastante amplia ya que es del 81,564% al 99,3342% en los diferentes departamentos del país.

La cobertura bruta, sin embargo, oscila entre 0,926231 y 1.148.539; aquí se cuenta con el primer cuartil en 1.020.565 y el tercer cuartil en 1.089.631, por otro lado una media en 918.545,33, una mediana en 1.064.933 y una desviación estándar en 37.634,44, con estos datos podemos observar que el mínimo es un dato atípico, y puede existir un sesgo ocasionado por este.

Por el lado de la deserción, tenemos que los datos oscilan entre 0,012632 y 0,053201, el primer cuartil se ubica en 0,0278 y el tercero en 0,03972, la media 0,033 y la mediana en 0,0325 y una desviación estándar de solo 0,01, con lo cual podemos decir que la deserción de los estudiantes entre 5 y 16 años es pequeña ya que esta entre el 1,263% y el 5,3201%.

A nivel de aprobación, los datos oscilan entre 0,858552 y 0,975008, el primer cuartil se ubica en 0,88597 y el tercero en 0,929093, se cuenta con una media de 0,907, una mediana de 0,905 y una desviación estándar de 0,027; es así como podemos observar que la gran mayoría de estudiantes aprueban su año escolar ya que este porcentaje se encuentra entre el 85,8552% y el 97,5008%.

A nivel de reprobación los datos oscilan entre 0,009037 y 0,096673, el primer cuartil se ubica en 0,05053 y el tercer cuartil en 0,075931, se cuenta con una media de 0,06 y una mediana de 0,0645, finalmente se tiene una desviación estándar de 0,022.

Finalizando con los atributos de la base de datos del gobierno, se tiene, que a nivel de repitencia, los datos oscilan entre el 0,007452 y el 0,036743, donde el primer cuartil se ubica en 0,01164 y el tercero en 0,031925, la media está en 0,02 y la mediana en 0,0153 y se cuenta con una desviación estándar de 0,01. Aquí podemos observar que el porcentaje de los estudiantes que repiten el año es bajo, inclusive más bajo que el de los que reprueban

el año, en este caso, los que repiten el año está entre el 0,7452% y el 3,6743% y los que lo reprueban está en 0,9037% y el 9,6673%.

Pasando a las cifras de la base de datos del DANE, el primer atributo p500 correspondiente a la cantidad de habitaciones con las que cuentan los hogares, tenemos que estos pueden variar entre 1 a 18 habitaciones, el primer cuartil se ubica en 3 habitaciones y el tercero en 4, se cuenta con una media de 3,3744, una mediana de 3 y una desviación estándar de 1,1974, para este caso se puede observar que las 18 habitaciones es un dato atípico y puede generar un sesgo entre los demás valores.

Para el atributo p5100, correspondiente al pago por la cuota de amortización del inmueble, se observa que los datos oscilan entre 98 y 930.000.000, donde el primer cuartil se ubica en 320.000 y el tercero en 1.000.000, con una media de 1.186.824,2024, una mediana de 560.000 y una desviación estándar de 12.923.946,9596, aquí se ve en caso donde las medidas se ven influenciados por los datos atípicos de los dos extremos, para este caso el 98 se considera que puede ser un valor no relacionado al valor a pagar, por el otro lado el 930.000.000 se considera asociado a una deuda muy grande y poco común.

Para el atributo p5130, correspondiente al valor estimado por el arriendo del inmueble, se observa que los datos oscilan entre 98 y 900.000.000, donde el primer cuartil se ubica en 200.000 y el tercero en 600.000, con una media de 583.207,7127, una mediana de 400.000 y una desviación estándar de 6.091.847,4498, aquí se ve un caso donde las medidas se ven influenciados por los datos atípicos de los dos extremos, para este caso el 98 se considera que puede ser un valor no relacionado al valor a pagar, por el otro lado el 900.000.000 se considera asociado a un arriendo bastante costoso y poco común.

Para el atributo p5140, correspondiente al valor real por el arriendo del inmueble, se observa que los datos oscilan entre 98 y 900.000.000, donde el primer cuartil se ubica en 290.000 y el tercero en 520.000, con una media de 464.740,8608, una mediana de 400.000 y una desviación estándar de 3.297.728,7054, aquí se ve un caso donde las medidas se ven influenciados por los datos atípicos de los dos extremos, para este caso el 98 se considera que puede ser un valor no relacionado al valor a pagar, por el otro lado el 900.000.000 se considera asociado a un arriendo bastante costoso y poco común.

Para el atributo nper, correspondiente a la cantidad de personas por hogar, esto puede variar entre 1 persona a 28, donde el primer cuartil se ubica en 2 personas y el tercero en 4 personas, para este caso la media es 3,2613, la mediana 3 y la desviación estándar 1,7888; en esta caso podemos ver como en la mayoría de casos en un hogar se llegan a presentar entre 1 persona a 4 personas; sin embargo, se presenta un dato atípico de 28 personas en un hogar, este podría estar ligado a una familia demasiado grande, hacinamiento dentro del lugar o un error a la hora de ingresar los datos.

Por otro lado, para el atributo ingtotugarr, correspondiente al ingreso en el hogar, se tiene que este oscila entre 0 y 158.000.000, donde el primer cuartil se ubica en 900.000 y el tercero en 227.000, la media se encuentra en 2.343.775,5191 y la mediana en 1.596.333,333 y se cuenta con una desviación estándar de 2.712.713,3747. Aquí se puede observar un dato atípico por el cual se pueden ver influenciados los datos, que es 158.000.000, sin embargo, el de 0 es más común debido a los niveles de pobreza e incluso de indigencia con la que se cuenta en el país.

Finalmente se tienen dos datos binarios que son: pobre e indigencia, los cuales nos indican si la persona entra en alguno de estos dos campos, es por esto que, no se puede entrar a analizar los datos a nivel de las diferentes medidas.

Al analizar los valores del sesgo (*skewness*) y la curtosis (*kurtosis*), se puede notar que algunos de los atributos presentan, con mucha seguridad, un sesgo en los datos o una presencia importante de *outliers*.

Este es el caso de las variables correspondientes a las preguntas **p5100**, **p5130** y **p5140**, es decir, el pago por cuota de amortización, el arriendo estimado del hogar (en caso de no estarlo pagando), y el arriendo real del hogar (en caso de vivir en arriendo) correspondientemente. En estas tres variables se observan valores del **sesgo** que se encuentran por fuera de lo común, lo que representa, de lejos, una alta asimetría en los datos. Al ser valores mayores a cero (al igual que el valor de este coeficiente en la mayoría de los atributos estudiados en este proyecto), se identifica que hay un sesgo positivo, es decir, al ser graficados los datos (en un histograma, por ejemplo) se observaría un sesgo a la derecha.

Asimismo, los coeficientes de **curtosis** en estas tres variables tampoco parecen comportarse cerca de valores esperados. Tomando como referencia, como es habitual en la estadística descriptiva, un valor de 3 para la definición de la distribución de los datos basado su comportamiento comparado con la curva normal es posible determinar que estos atributos (así como la mayoría de los atributos estudiados en este proyecto) presentan en sus datos una distribución **leptocúrtica**, lo que representa, en otras palabras, una distribución con una “empinación” mayor a la de la curva normal.

Al presentar valores tan altos en este coeficiente, es posible indicar que los datos cuentan con presencia de valores atípicos (u *outliers*), y que estos tienen una gran influencia en la cola de la distribución, haciéndolas distribuciones de cola pesada (***heavy-tailed distribution***).

Los valores presentados en ambos coeficientes ponen en duda la calidad de los datos recogidos en las tres variables mencionadas, por lo cual se requiere de una revisión exhaustiva de los datos contenidos en estas para identificar problemas (como por ejemplo los valores mínimos expuestos en la tabla de estadísticas descriptivas, en donde vemos valores de 98 y 9, los cuales no tienen sentido para el contexto manejado). Asimismo, es necesario, después de este análisis exploratorio, realizar una limpieza de los datos para llegas a una vista minable que no se vea afectada por los problemas que indican los coeficientes mencionados.

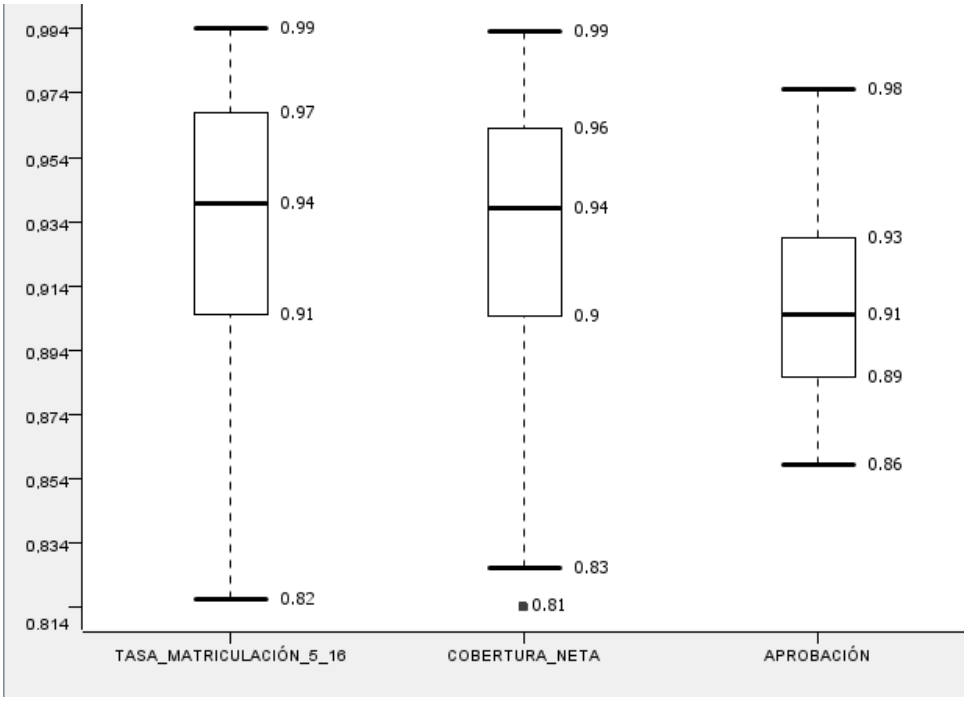
Otra de las variables con valores inusuales en estos coeficientes es **ingtotugarr**, que corresponde al ingreso total de la unidad de gasto con imputación de arriendo a propietarios y usufructuarios. Esta variable presenta un valor en el coeficiente de sesgo de 6,13 y en el de curtosis uno de 109,3. En el caso del **sesgo**, se determina que los datos tienen una distribución considerablemente alta, siendo además una distribución sesgada a la derecha. Esto se puede explicar teniendo en cuenta que se está hablando de ingresos de los hogares, en donde comúnmente encontramos este tipo de sesgos, siendo que gran parte

de la población se concentran a la izquierda de la distribución (valores más bajos o incluso muy cercanos a cero), haciendo que la cola de la distribución sea más larga a la derecha.

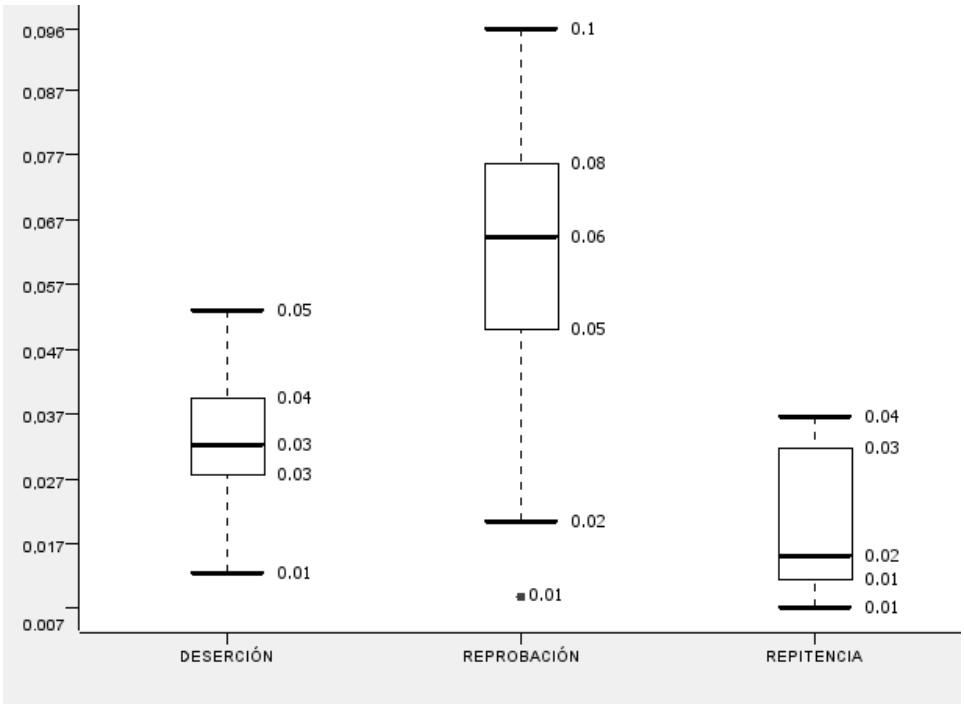
Tomando ahora la **curtosis**, es posible observar la fuerte presencia de valores atípicos en los datos, haciendo de la distribución, una distribución leptocúrtica. Esto se puede explicar porque, mientras gran parte de los datos se concentran en valores bajos (teniendo incluso una distribución con un tercer cuartil en solo $2'827.000), podemos encontrar puntos de datos que representan, atípicamente, ingresos de hasta $158'000.000. Estos no necesariamente deben representar datos erróneos, sino que puede dar muestra de datos de hogares con ingresos muy superiores, perteneciendo quizá a hogares de empresarios o políticos.

1.4. Visualización y análisis de métricas de dispersión

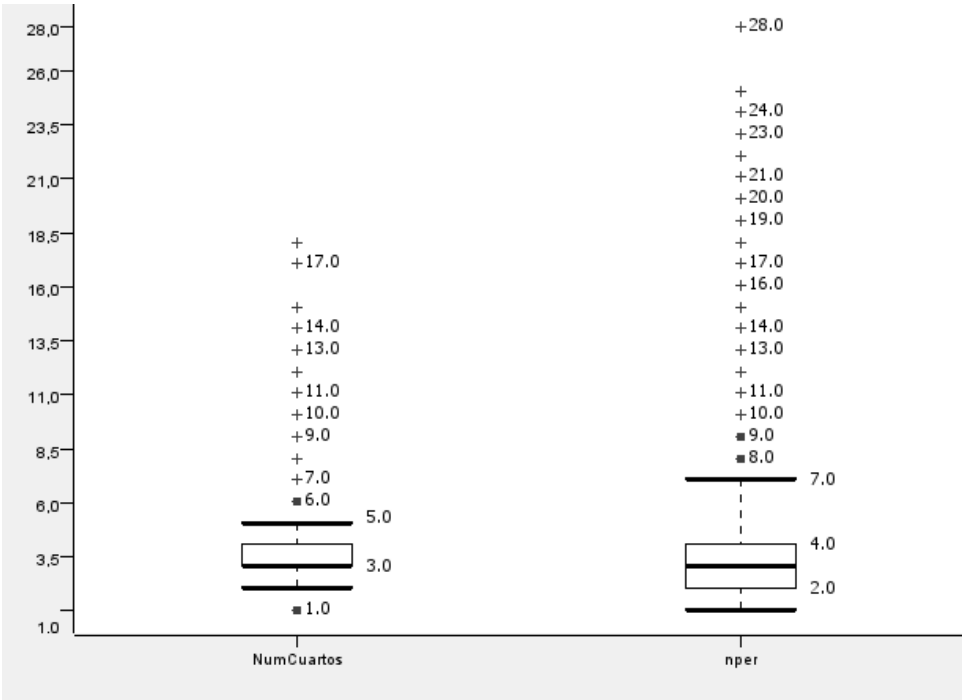
Después de la carga y análisis de la información mencionada, continuamos con la implementación de 3 nodos de visualización de gráficos, con el fin de analizar las métricas de dispersión de las variables en cuestión. Iniciamos con el nodo "Color Manager" el cual nos permite definir los colores de las visualizaciones para los gráficos que vamos a trabajar y poder diferenciar las variables. Continuamos insertando los nodos "Box Plot (local)", "Bar Chart" y "Scatter Plot (local)" evidenciando los resultados a continuación:

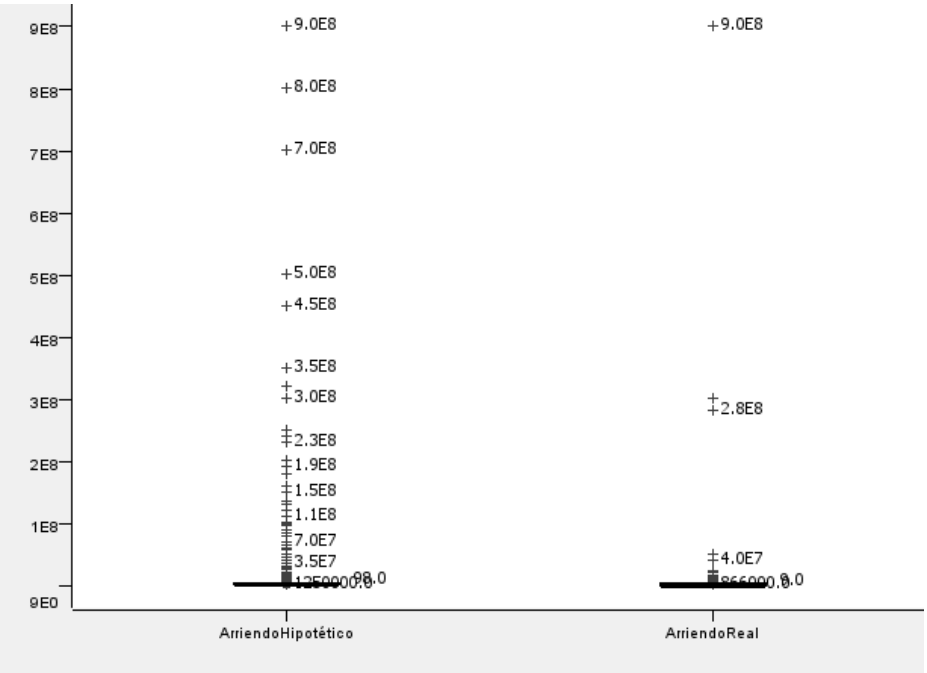


Mediante el gráfico Box Plot o de Cajas y Bigotes, para las variables Tasa de Matriculación, Cobertura Neta y Aprobación, correspondientes a la base de datos de educación, vemos una presencia casi nula de datos atípicos. Siendo estas tasas en rangos entre 0.8 y 1 con medianas superiores a 0.9 indica resultados positivos en el país para el año 2019 en cuanto a estas variables.

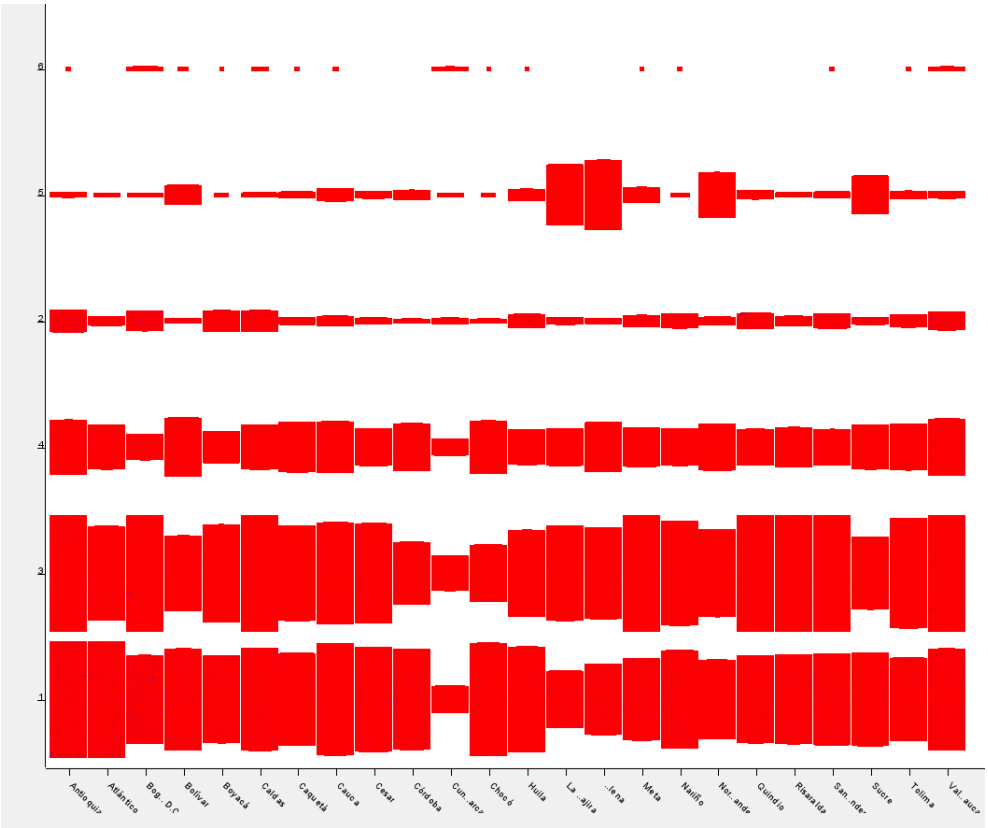


En el caso de las variables Deserción, Reprobación y Repitencia podemos ver, de manera congruente con lo analizado anteriormente, Box Plot con rangos entre 0.01 y 0.1 indicando resultados positivos para Colombia en el 2019, sim embargo cabe resaltar la variable Reprobación con una mediana de 0.06 siendo la más alta entre las variables en cuestión.

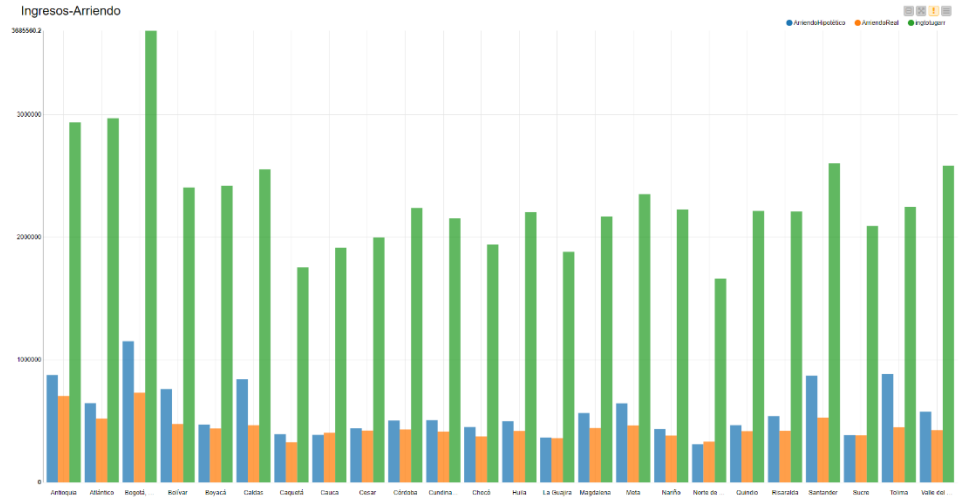




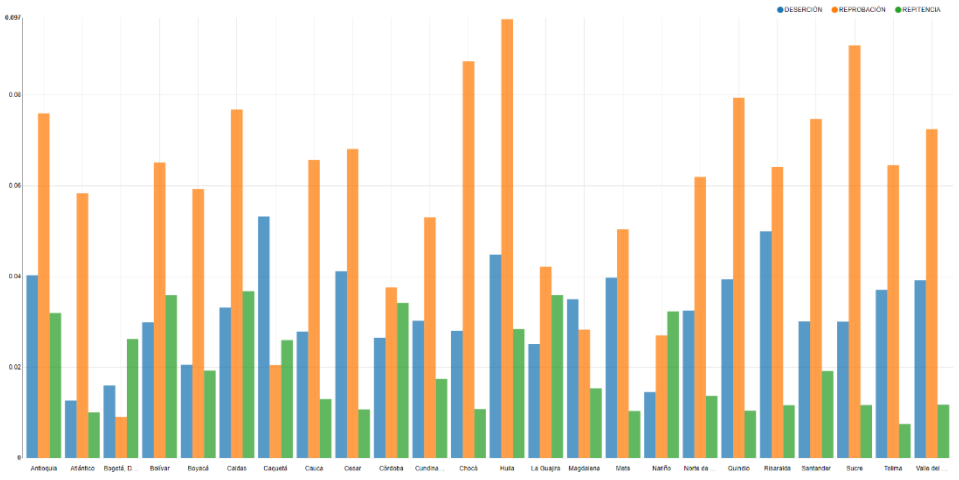
Continuando con la base de datos del Dane, realizando el Box Plot para las variables NumCuartos, nper, ArriendoHipotético y ArriendoReal encontramos una cantidad considerable de datos atípicos fuera de las cajas de cada variable, estos serán tomados en cuenta en la etapa de limpieza y tratamiento de datos, con el fin de generar información más acertada con un menor impacto de este tipo de datos en los resultados esperados.



Utilizando el gráfico Scatter Plot (local), con el fin de desarrollar una matriz de dispersión, relacionamos las variables Departamento y TipoVivienda, encontrando el mayor volumen en los tipos de vivienda 1 (Propia, totalmente pagada) y 3 (En arriendo o subarriendo) con una tendencia muy parecida por departamento, continuando con el tipo 4 (En usufructo).



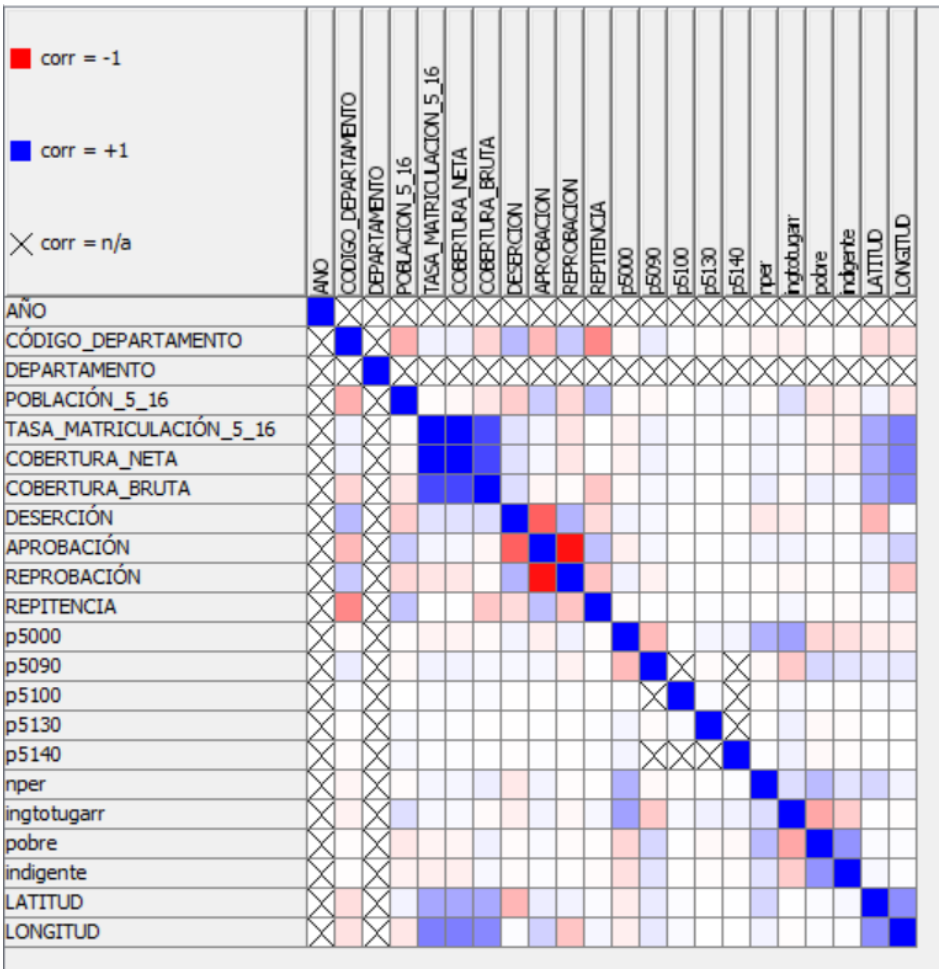
Mediante el gráfico “Bar Chart” comparamos las variables de Ingtotugarr junto con ArriendoHipotético y ArriendoReal en la cual podemos ver a Bogotá con los valores más altos para las tres variables versus el resto de los departamentos, de igual forma evidenciamos una tendencia parecida para las tres variables por departamento, a medida que disminuye el ingreso de igual forma lo hace el arriendo.



Con el fin de identificar relaciones entre las variables en esta primera etapa del proyecto, utilizamos nuevamente el gráfico de barras, para ver las variables de Deserción, Reprobación y Repitencia por departamentos. Comparando con el gráfico anterior podemos ver que algunos de los departamentos con ingresos más bajos tienen unas de las tasas más altas de Reprobación siendo un dato muy interesante el cual será analizado más adelante.

1.5. Análisis de correlación:

Gráfico de correlación lineal:



Después analizar los gráficos de correlación, se detectaron variables con coeficientes de correlación que indican que existe una relación fuerte entre variables de la base de datos, estas variables serán tratadas en la etapa de limpieza de datos.

Estas variables son:

- Cobertura Neta / Tasa de Matriculación
- Cobertura Bruta / Tasa de Matriculación
- Reprobación / Aprobación

S First column name	S Second column name	D Correla...	D p value	I Degree...
TASA_MATRICULACIÓN_5_16	COBERTURA_NETA	0.99975854...	0.0	231829
TASA_MATRICULACIÓN_5_16	COBERTURA_BRUTA	0.72084555...	0.0	231829
APROBACIÓN	REPROBACIÓN	-0.9317357...	0.0	231829

Etapa 2: Limpieza de datos:

Teniendo en cuenta las conclusiones que se pueden obtener a partir del análisis exploratorio de datos expuesto previamente en este documento, es necesario realizar un proceso de limpieza de datos para poder obtener resultados en nuestro análisis a partir de visualizaciones y posteriormente realizar una vista minable que no se vea afectada por los problemas típicos que presentan los datos al momento de su extracción.

2.1 Valores Únicos o Distintos:

Inicialmente, y en base al análisis que se quiso desarrollar en este proyecto, se hizo un primer filtro en la variable correspondiente al año tanto para la base del ministerio de educación como para las bases provistas por el DANE. En este filtro, que se discutió al momento de definir el enfoque del proyecto, se tomó la decisión de analizar los datos del último año disponible en ambas bases, es decir, el 2019.

Habiendo realizado este filtro, la variable año contenía ahora un único elemento en todos sus registros (2019), es por esto por lo que se tomó la decisión de eliminar la variable del proceso de preparación de los datos.

2.2 Valores Faltantes:

Dos variables que cabe mencionar en este apartado son las correspondientes a las preguntas p5130 y p5140, que representa el arriendo real pagado por el hogar y el arriendo hipotético (es decir, cuánto pagarían si tuvieran que pagar arriendo, pregunta que responden únicamente aquellos hogares que no pagan arriendo ya que, por ejemplo, poseen casa propia). En estas dos variables, debido a ser una pregunta sujeta a la existencia o no de un pago de arriendo mensual para el hogar, tenía un valor nulo en cada registro en que la otra columna tuviera un valor no nulo. Es por esto por lo que, para tratar estos datos, se creó una variable derivada que permitiera solucionar esto. Las variables derivadas en este trabajo se explicarán en la etapa de preparación de la vista minable.

2.3. Atributos con valores Atípicos:

Al momento de realizar la exploración de los datos faltantes en los atributos del estudio, fue posible notar que uno de los atributos, en específico la pregunta 5100 de la base de datos de medición de pobreza monetaria correspondiente a la cuota de amortización mensual pagada por el hogar encuestado, estaba representada en más del 96% de sus registros por datos nulos. Es por esto por lo que se tomó la decisión de dejar esta variable por fuera del estudio.

2.4. Registros Atípicos:

En tres de las variables del conjunto de datos de pobreza monetaria del DANE se encontraron valores atípicos:

Los dos primeros son la variable correspondiente al pago mensual de arriendo y la del pago mensual hipotético de arriendo. En estas se encontraron valores sumamente altos, así como sumamente bajos que no tenían ningún sentido para el contexto manejado. Para solucionar este inconveniente se determina un rango para este atributo de entre $10.000 y $15.000.000, los demás registros no se tienen en cuenta para el ejercicio.

Por otro lado, se hizo un tratamiento similar a la variable correspondiente al ingreso del hogar encuestado. En este caso, se definió un rango que va desde $0 hasta $50.000.000, dejando por fuera 51 filas (Menos del 0.1% de los registros).

2.5. Atributos Redundantes:

Tal como se observó en la etapa de exploración de datos, al revisar las correlaciones entre los atributos, se puede determinar que algunos de ellos cuentas con unas correlaciones altas entre sí, siendo este el caso de tres de las variables correspondientes a las tasas de educación brindadas por la base de estadísticas del Ministerio de Educación Nacional. Para este caso, se determinó cuáles de estas tasas presentan mejor interpretabilidad al momento de realizar los análisis posteriores. Es por esto por lo que se tomó la decisión de eliminar las variables correspondientes a la cobertura bruta, la cobertura neta y la reprobación de cada departamento, dejando entonces la aprobación y la tasa de matriculación.

Etapa 3: Creación de Vista Minable:

3.1. Generación de variables tipo 1 y 2

Continuando con la tercera etapa del proyecto de creación de vista minable con el objetivo de organizar los datos lo mejor posible para reducir y facilitar el trabajo que debe hacer el algoritmo de minería, cada paso fue almacenado en un metanodo con el fin de tener el programa lo más ordenado posible. Comenzamos con la generación de variables derivadas tipo 1 y 2.

3.1.1. Variables derivadas tipo 1

Revisando las variables ArriendoHipotético y ArriendoReal encontramos que podían unirse debido a que si una tenía valor positivo la otra aparecía con valor = 0 por lo tanto, mediante el Nodo “Column Expressions” con el fin de mantener la información de que tipo de arriendo corresponde con las siguientes indicaciones:

```
if (column("ArriendoHipotético") > 0)
{0}
else if (column("ArriendoReal") > 0)
{1}
```

Continuamos con el nodo “Column Agregator” el cual nos permite realizar la suma de las dos variables seleccionándolas, indicando la operación SUM y generando una nueva columna con el fin de remover las dos variables mencionadas. Seguido de un “RowFilter” con el fin de filtrar en un rango entre 10.000 y 15.000.000 que permita reducir los valores atípicos aparentemente por errores en digitación con supuestos arriendos de 98 o incluso de 900.000.000. Con el fin de resumir hasta la limpieza de datos y creación de esta variable creamos un nodo “Group By” con el objetivo de tabla dinámica y un “Bar Chart” como gráfico de visualización, para generar y analizar información importante después de su debido proceso de limpieza y poder incluirla en la infografía.

Creamos la variable PorcentajeArriendo mediante un nodo “Math Formula” con la siguiente indicación $Arriendo$/$ingtotugarr$ la cuál decidimos implementar con el objetivo de identificar el % de ingresos que la población utiliza para el pago de sus arriendos. Continuamos con un nodo “Column Expressions” con el fin de validar que la variable PorcentajeArriendo no contenga errores

como una división sobre valor 0, esto se pudo determinar analizando que si una persona realmente paga un arriendo debe tener un ingreso superior a este. mediante el siguiente código:

```
if (column("PorcentajeArriendo") > 1 && column("PagoArriendo") == 1)
{"error"} else
{"ok"}
```

Seguido de esto realizamos un “Row Filter” con el objetivo de filtrar las filas que contengan error de acuerdo a los resultados del nodo anterior y evitar de igual forma cualquier valor que afecte los resultados finales.

Variables derivadas tipo 2

Creamos la variable tipo 2 DistanciaMediaIng la cual hace referencia a la distancia del ingreso de su media a partir del nodo “Math Formula” con la siguiente indicación: $ingtotugarr$-COL_MEDIAN($ingtotugarr$) esta variable queda disponible para futuros análisis sobre los temas observados.

3.2. Numerización 1 a N:

Continuamos con la Numerización aplicada a las variables pobre e indigente con el objetivo de unificar estas columnas en una llamada NivelPobreza mediante el nodo “Column Expressions” con la siguiente indicación:

```
if (column("pobre")== 1&& column("indigente")==1)

{1}

else if(column("pobre")== 1&& column("indigente")==0)

{2} else {3}
```

Seguido de un nodo “Column Filter” para filtrar las variables pobre e indigente.

3.3. Numerización 1 a N:

En el metanodo de normalización utilizamos el nodo “Normalizer” mediante el método de Mínimos y Máximos definidos entre 0 y 1 para las variables Poblacion, NumCuartos, nper, Arriendo, DistanciaMediaIng y NivelPobreza.

Más adelante realizamos una transformación mediante el nodo “Math Formula” para las variables Arriendo, PorcentajeaArriendo y DistanciaMediaIng incluidas a partir de la operación log($variable$) incluidas en el metanodo Transformación.

3.4. Discretización:

Para finalizar, en el metanodo de la discretización utilizamos el nodo “Numeric Binner” con el cual discretizamos, de acuerdo a sus cuartiles en las estadísticas, las variables de Ingreso y Población, siendo MuyBajo, Bajo, Medio, Alto, Muy Alto y Bajo, Medio, Alto respectivamente.

Etapa 4: Infografía

Dado que la infografía es una herramienta de comunicación para presentar de manera sencilla, resultados que reflejan realidades como esta, es pertinente mencionar que se quiso destacar a manera de resumen y las estadísticas de educación departamental presentadas por el gobierno para el año 2019. Inicialmente se planteó mostrar datos relevantes de cada uno de los sets, para posteriormente mostrar relaciones entre ellos. Para esto, se realizó primero la carga de datos y una limpieza de estos, con el fin de descartar datos atípicos.

Por un lado se presentan los tres departamentos donde sus poblaciones tienen en promedio los ingresos más bajos del país, se inicia con el Chocó, se pasa a Norte de Santander y finaliza con la Guajira. Por otro lado se tiene el opuesto de estos datos donde se presentan los tres lugares con ingresos más altos del país, donde se inicia con Bogotá, se pasa a Antioquia y se finaliza con Atlántico. Si nos vamos a conocimientos populares, podemos observar que estos datos concuerdan con lo que comúnmente se conoce, donde se sabe que en ciudades como Bogotá, Medellín y Barranquilla, viven empresarios de gran calibre y existen empresas multinacionales, en ambos casos estos datos mueven los ingresos promedio a ser más altos, por otro lado, siempre se ha dicho que departamentos como el Chocó y la Guajira son de los más pobres del país y no se tienen tantas oportunidades para generar ingresos, adicionalmente en departamentos como la Guajira se tiene una alta presencia indígena y estos no presentan ingresos.

Posteriormente se pasa a destacar algunos datos del nivel educacional, por un lado se tiene, que de la población que estudió, el 10% reprueba el año y solo el 2% lo repite, esto lleva a pensar en diferentes factores que pueden influir en que estos números no sean similares, por un lado se puede considerar que algunos estudiantes que pierden el año deciden abandonar sus estudios, otros pueden perder el año pasarse a otro colegio y presentar un examen para ser promovidos. Adicionalmente cabe resaltar que no todas las repitencias van ligadas a la perdida del año, algunas ocurren porque el estudiante decidió que no se sentía lo suficientemente preparado para afrontar el siguiente año escolar y decide repetir u otro caso en cuando el estudiante se cambia de colegio y al presentar el examen de admisión, en la nueva institución, no alcanza el puntaje para el grado al que quiere entrar y lo hacen repetir el curso; sin embargo cabe resaltar que ambos casos presentados son datos atípicos.

Siguiendo la línea de los datos de escolaridad, se presentan los departamento con las tasas de matriculación más bajas, por un lado se tiene a Caldas, donde del 100% de la población que puede ser matriculada, solo el 86% se matricula, por otro lado se tiene a Cauca, donde solo el 85% se matricula, seguido de Nariño con el 84% de personas matriculadas, se sigue con el Chocó, donde solo el 83% se matricula y se finaliza con el Valle del Cauca, donde solo se matricula el 82% de la población en edad de asistir al colegio. A pesar de que estos son los departamentos con menor tasa de matriculación, que todos tengan porcentajes por encima del 80% es sumamente alentador, ya que esto nos dice que cada vez son más los jóvenes y niños que pueden acceder a este derecho de la educación.

Adicionalmente se pasa a realizar una comparación entre el promedio de arriendos y el promedio de ingresos por departamentos, con el fin de denotar el porcentaje que en promedio se destina para cubrir gastos de arriendo de vivienda. Donde se destacan los

departamentos con mayores ingresos y arriendos vs los que tienen menores ingresos y costos de arriendo. Se inicia con Bogotá el cual tiene un ingreso promedio de más de 3 millones de pesos y un costo en arriendo de más de 808 mil pesos, se continua con Antioquia donde los ingresos llegan casi a los 3 millones de pesos y en arriendo se gasta más de 600 mil pesos, con valores muy similares se sigue con Atlántico, pasando a los departamentos de Santander y Valle del Cauca, los cuales presentan datos similares, a nivel de ingresos se tiene que en promedio es superior a los 2 millones de pesos y en arriendo más de 400 mil pesos. Por el otro lado a nivel de los departamentos con menores ingresos, como se menciono anteriormente, se encuentran Cauca, con más de 1 millón de pesos en ingreso promedio y de arriendo más de 400 mil pesos, se sigue con Chocó con un ingreso promedio de más de 1 millón de pesos y un costo promedio de arriendo de más de 300 mil pesos, se continua con la Guajira con datos muy similares a los del Chocó, al igual que Caquetá y Norte de Santander. Aquí se puede observar que de forma constante, los costos de arriendo equivalen a solo el 20% de los ingresos.

Posteriormente se presenta una tabla con el top 9 de los departamentos con ingresos más bajos y el top 9 con las peores tasas de educación, adicionalmente en esta tabla se presentan los porcentajes de aprobación, repitencia y deserción, cabe destacar que en los departamentos donde el valor en alguno de estos atributos es '-' no significa que no se tenga el valor, si no que este no corresponde a los valores más bajos de dicho atributo.

Finalmente se destacan algunas de las conclusiones a las que se llegaron a lo largo del proyecto.

Etapa 5: Conclusiones

Se pudo observar que la tasa de matriculación en los diferentes departamentos es superior al 80% en todos los casos, y la tasa de deserción, a su vez, es bastante baja (alrededor del 2 %) lo cual nos indica que existe una alta cobertura de educación básica y media, permitiendo que bastantes niños y jovenes disfruten del derecho a la educación.

Los ingresos en Colombia varían por departamento afectando la distribución de estos en las necesidades básicas del hogar.

Se evidencia una clara relación entre bajas tasas de ingresos y bajas tasas de acceso a la educación (matriculación).

El gasto de los hogares en arriendo es proporcional al ingreso, con un promedio del 20% de sus ingresos destinados a este concepto.