Examen

Minerva Márquez Castillo

25 de Mayo del 2020

1 PARTE A

1. Para cada AGEB del Municipio de Toluca, estima cuántos bebés de 0 a 6 meses de edad habitan ahí el día de hoy. Explica tu razonamiento en menos de 300 palabras. En lista tus fuentes y presenta los resultados. (Hint: revisa el CPV 2010 , y el Inventario Nacional de Vivienda 2016). Puedes usar cualquier otra fuente que consideres relevantes.

   En la base de datos "$CPV2010$" se utilizó la columna de la Población de niños de 0 a 2 años. Supongamos una distribución uniforme en las edades de 0 a 2 años.

$$f(x) = \begin{cases} \frac{1}{2} & \text{si } 0 \leq x \leq 2 \\ 0 & \text{si } 0 < x \quad o \quad x < 2, \end{cases} \tag{1}$$

   En el código R se trabajo con la base de $CPV2010$ y se utilizó la columna de la Población Total de 0 a 2 años, y tomando la suma de cada AGEB se estimo la cantidad de niños entre 0 a 6 años. Se multiplico por un $\frac{1}{4}$ a la suma de cada uno de los AGEB del municipio de Toluca para obtener una cuarta parte de la población de 0 a 2 años son bebés de 0 a 6 años. Algunos datos de los bebés de 0 a 6 meses, como se muestra en la siguiente tabla 3:

AGEB	0027	002A	0031	0034	0041
BEBÉS	112.25	70.75	53.50	41.50	44.50
AGEB	0050	0053	0056	0065	0068
BEBÉS	29.00	33.50	32.25	27.75	74.00
AGEB	0072	0075	007A	0084	0091
BEBÉS	140.75	7.00	44.25	56.00	8.75

Table 1: Datos de los niños de 0 a 6 años del municipio de Toluca.

2 PARTE B

Descarga la Base de datos histórica de Quién es Quién en los Precios de Profeco y resuelve los siguientes incisos. Para el procesamiento de los datos y el análisis

exploratorio debes debes usar Spark SQL en el lenguaje de programación de tu elección.

En el código R usando Spark SQL, se trabajo con las columnas categoricas, estado, cadena comercial, precio y producto. En los programas se explica a detalle el uso de estás variables.

1. Procesamiento de los datos

   (a) ¿Cuántos registros hay? Número de registros 62,530,715

   (b) ¿Cuántas categorías? Son 40 categorias en relaidad ya que hay datos no observados

   (c) ¿Cuántas cadenas comerciales están siendo monitoreadas? Son 704 categorias ya que hay un renglón vacio

   (d) ¿Cómo podrías determinar la calidad de los datos? ¿Detectaste algún tipo de inconsistencia o error en la fuente? La calidad de los datos es realizar el proceso de ETL, algunas transformaciones comunes son: Seleccionar sólo ciertas columnas para su carga (por ejemplo, que las columnas con valores nulos no se carguen), traducir códigos, validar los datos, etc.

   Son datos estructurados y pueden corromperse o puede haber fallas en la captura de los mismos, como es en este caso que hay espacios vacíos en la columna y también hay nombres que no son estado.

   (e) ¿Cuáles son los productos más monitoreados en cada entidad?

Estado	Producto	Venta
AGUASCALIENTES	LIMPIADOR LIQUIDO P/PISO	5167
BAJA CALIFORNIA	JABON DE PASTA	4164
BAJA CALIFORNIA SUR	SALSA DE SOYA	1228
CAMPECHE	CARNE RES	4354
CHIAPAS	PIÑA EN ALMIBAR	2392
CHIHUAHUA	LAVADORAS	11479
COAHUILA DE ZARAGOZA	TOALLITA HUMEDA LIMPIADORA	6053

Table 2: Algunos estados con sus productos más vendidos

   (f) ¿Cuál es la cadena comercial con mayor variedad de productos monitoreados? La tabla es 4

2. Análisis exploratorio

   (a) Genera una canasta de productos básicos que te permita comparar los precios geográfica y temporalmente. Justifica tu elección y procedimiento

   (b) ¿Cuál es la ciudad más cara del país? ¿Cuál es la más barata?

Comercializadora	Producto	Venta
7 ELEVEN	RASTRILLOS DESECHABLES	501
ABARROTES SUPER CABRERA	TORTILLA DE HARINA DE TRIGO	226
ADOSA	REGLAS	29
ALMACENES ZARAGOZA	NOPAL	125
ALSUPER	MICARDIS PLUS	2313
ANDA FARMACIAS	BENZETACIL	16
AUTOSERVICIO ZARAGOZA	SHAMPOO	286
BENAVIDES	NORVAS	1280
hline BODEGA AURRERA	SILDENAFIL	2313

Table 3: Algunas cadenas comerciales con sus productos mas vendidos.

(c) ¿Hay algún patrón estacional entre años?

(d) ¿Cuál es el estado más caro y en qué mes?

(e) ¿Cuáles son los principales riesgos de hacer análisis de series de tiempo con estos datos?

3. Visualización

(a) Genera un mapa que nos permita identificar la oferta de categorías en la zona metropolitana de León Guanajuato y el nivel de precios en cada una de ellas. Se darán puntos extra si el mapa es interactivo

3 PARTE C

El documento del caso en resumen habla del efecto que tiene la iniciativa BOPS (Buy Online Pickup at Store) respecto a las ventas online o por $B\&M$ (en tienda) de Home and Kitchen. Ésta inciativa se implementó en Octubre de 2011 en las 67 tiendas de Estados Unidos, las 17 tiendas de Canadá aún no estaban listas. Se quiere indagar si en verdad ésta inciativa aporta beneficio a las ventas.

Según las bases de datos, tenemos registros de cada venta en tiendas locales y online por semana desde abril de 2011 a abril de 2012. Cada observación corresponde a una venta identificada por el id de la tienda ó a una venta realizada online en una cierta unidad de DMA (Designated Market Areas), que es una unidad geográfica estándar .

Proponemos el modelo de regresión

$$y = \alpha_1 x_1 + \alpha_2 x_2 + \beta x_3 + \gamma x_2 x_3$$

en el que cada variable tiene como valores e interpretación:

1. $x_1 = 0$, si la venta fue en tienda física y $x_1 = 1$ en otro caso.

2. $x_2 = 1$, si la venta se ejecutó en línea y $x_2 = 0$ en otro caso. En otras palabras $x_2 = 1 - x_1$.

3. $x_3 = 1$ si la venta se realizó antes de la implementación de BOPS y $x_3 = 0$ si fue después.

Definimos a cada unidad de observación a la cantidad total de ventas que se hicieron por todas las tiendas físicas (o ventas en línea según sea el caso) entre semana. Es decir, no habrá distinción de qué tienda física realizó tal venta ni en qué región se realizó si fue en línea. Esto nos permite encontrar el efecto que tiene la variable γ del modelo.

En las lineas de CÓDIGO DE R se definen solo dos variables, pero en la función lm() se deseglosan las variables del modelo. Ajustando el modelo en R, obtenemos la siguiente información:

coeficientes	valor
α_1	4,545,581
α_2	3,094,
β	-454,633
γ	33,886

Table 4: Valores de la constantes α_1, α_2, β y γ

Interpretación de los coeficientes con α_1 por cada unidad de venta realizada en físico aumenta, y para α_2 por cada unidad de venta realizada en línea aumenta y β por unidad realizada después de haber implementado BOPS. El coeficiente γ el efecto que adiciona el haber implementado BOPS y que la venta se haya realizado por línea.

Se obtuvieron otros datos del modelo

coeficientes	Error estándar	t value	$Pr(>\|t\|)$
α_1	182943	24.847	$< 2e-16 * **$
α_2	186427	16.601	$< 2e-16 * **$
β	258720	-1.757	0.0819
γ	371263	0.091	0.9275

Table 5: Interpretación del modelo.

En la tabla 5 se observa que los coeficientes α_1 y α_2 son variables significativas y que el valor de $R = 0.9379$ está cerca de uno lo que significa que es un buen modelo.

3.1 Preguntas del pdf examen

1. ¿Deberían expandirse a canadá? Sí, esperando los mismos resultados que pasó en Estados Unidos. La aportación que hace la iniciativa BOPS es aumentar las ventas en casi $33,886$ unidades

2. ¿Cuántos millones de dólares se ganaron o perdieron a partir del programa? Se perdieron $454,633$ por semana según la intepretación mencionada antes de β

En el link para acceder a los programas realizados en R https://github.com/Minerva0401/Solucion-Opi-Analytics