Solución Examen Opi Analytics

Eduardo Santiago Hernández

May 16, 2021

1. Ejercicio A.

   (a) ¿Qué pruebas identificarías para asegurar la calidad de estos datos? No es necesario hacerlas. Sólo describe la prueba y qué te dice cada una.

   - La disponibilidad de los datos: lo que se refiere el formato por el cual se encuentran
   - La metodología con la que se diseñó la base de datos: lo que se refiere el muestreo, el diccionario de las variables , entre otros.
   - L cantidad de nulos: lo que se refiere a identificar los renglones que carecen de información.
   - valores no pérmitidos registrados en una variable. lo que indicaría en hacer un resumen de las variables.

   (b) ¿Cuántos delitos registrados hay en la tabla? ¿Qué rango de tiempo consideran los datos?

   Hay $808,871$ y el rango de tiempo abarca desde el mes de Junio de 1906 al mes de Junio de 2019.

   (c) ¿Cómo se distribuye el número de delitos en la CDMX? ¿Cuáles son los 5 delitos más frecuentes?

   Se muestran los 5 delitos más frecuentes

```
VIOLENCIA FAMILIAR
                         69517
             ROBO DE OBJETOS
                         52214
ROBO A NEGOCIO SIN VIOLENCIA
                         51426
                        FRAUDE
                         45349
           DENUNCIA DE HECHOS
                         44433
```

   (d) Identifica los delitos que van a la alza y a la baja en la CDMX en el último año (ten cuidado con los delitos con pocas ocurrencias).

(e) ¿Cuál es la alcaldía que más delitos tiene y cuál es la que menos?.¿Por qué crees que sea esto?

La que más tiene es la alcaldía CUAUHTEMOC con $131,397$ delitos registrados, por otro lado existen registros de alcaldías con un solo registro, por ejemplo:

```
# ABALA
# 1
# ACAMBARO
# 1
# ACAXOCHITLAN
# 1
# ACONCHI
# 1
# ACTOPAN
# 1
# ACULCO
# 1
# AGUA DULCE
# 1
# AGUA PRIETA
# 1
...
```

Lo que no necesariament significa que estas últimas alcaldías tengan poca delincuencia, pudiera ser que el segimiento de estas mismas no se esté dando.

(f) Dentro de cada alcaldía, cuáles son las tres colonias con más delitos

Mostramos el *head* de la consulta en SQL.

	alcaldia_hechos	colonia_hechos	numberDelitos	delitos_rank
1			1008	1
2		SIN INFORMACIÓN	3	2
3	ABALA		1	1
4	ABASOLO		2	1
5	ACAMBARO		1	1
6	ACAMBAY		3	1

Podemos observar que hay registros de los cuales no se sabe en qué colonia y alcaldía ocurrieron los hechos.

(g) ¿Existe alguna tendencia estacional en la ocurrencia de delitos (mes, semana, día de la semana, quincenas)?

(h) ¿Cuáles son los delitos que más caracterizan a cada alcaldía? Es decir, delitos que suceden con mayor frecuencia en una alcaldía y con menor frecuencia en las demás.

2. Ejercicio B.

Se utiliza la libreria pyspark.
Procesamiento de los datos

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

(b) ¿Cuántas categorías? hay 42

(c) ¿Cuántas cadenas comerciales están siendo monitoreadas? son 706

(d) ¿Cómo podrías determinar la calidad de los datos? Identificando la cantidad de valores nulos por cada columna, así como la cantidad de renglones que contienen al menos un nulo.
Por otro lado, tomando los valores distintos en variables categóricas (por ejemplo la variable estado) e identificar si realmente son los 32 entidades, así como un conteo de los municipios de cada estado y hacer una comparativa con datos del Inegi.

(e) ¿Detectaste algún tipo de inconsistencia o error en la fuente?
Al explorar con algunas consultas sql, pude notar que hay datos nulos en algunas columnas, asi como un mal capturamiento de los datos en la variable estado, pues existen colonias registradas en esa variable

(f) ¿Cuáles son los productos más monitoreados en cada entidad? Mostramos dos registros de cada entidad, donde es notable ver que en la variable estado se registraron datos que no son los correspondientes de un estado.

estado	producto	count	rank
QUINTANA ROO	FUD	34846	1
QUINTANA ROO	REFRESCO	34367	2
NUEVO LEÓN	DETERGENTE P/ROPA	50307	1
NUEVO LEÓN	REFRESCO	49592	2
SINALOA	REFRESCO	33115	1
SINALOA	DETERGENTE P/ROPA	27177	2
TABASCO	REFRESCO	28754	1
TABASCO	DETERGENTE P/ROPA	26431	2
BAJA CALIFORNIA	REFRESCO	37243	1
BAJA CALIFORNIA	DETERGENTE P/ROPA	23395	2
TLAXCALA	REFRESCO	43904	1
TLAXCALA	DETERGENTE P/ROPA	41398	2
COAHUILA DE ZARAGOZA	FUD	28613	1
COAHUILA DE ZARAGOZA	REFRESCO	26889	2
null	LECHE ULTRAPASTEU...	804	1
null	REFRESCO	553	2
ESQ. SUR 125"	PAN BLANCO BOLILLO	130	1
ESQ. SUR 125"	TORTILLA DE MAIZ	2	2
COL. EDUARDO GUERRA	REFRESCO	275	1
COL. EDUARDO GUERRA	JABON DE TOCADOR	270	2

only showing top 20 rows

(g) ¿Cuál es la cadena comercial con mayor variedad de productos monitoreados? es Soriana con $1,059$ productos.

3. Ejercicio C.

De acuerdo a lo que se lee en el caso, habla sobre el crecimiento de ventas en línea que habrá durante los años 2012 en adelante, en general, sin importar el tipo de empresa que sea.

Home and Kitchen es una compañía que empezó sus ventas en línea por el año 2007 y en aquel entonces, quien encabezaba la división de ventas en línea, Kristen Schwarz, lanzó una iniciativa que consistía en vender productos en línea con la opción de que el cliente podría recogerlo en una tienda cercana, llamada *BOPS* por su abreviación en ingles. La iniciativa se registro durante un periodo de ventas por semana durante 6 meses antes y después de *BOPS*. Los datos mostrados en el pdf no parecen dar crédito a la idea de Schwarz.

Para indagar el problema, proponemos un modelo de diseño factorial, con interacciones, donde se tienen dos factores con dos niveles, el factor A da indicio de si la venta registrada fue antes o después de la iniciativa y el factor B da indicio de si la venta fue realizada directamente en una tienda física o fue en línea lo que nos da un total de $2 \times 2 = 4$ posibles combinaciones.

Hacemos el filtrado de los datos considerando solo aquellos registros de ventas realizadas dentro de Estados Unidos ó si fue una venta en línea cuyo cliente recogió el producto en una tienda cercana a no más de 50 millas.

Sea y el total de ventas realizadas de la compañía durante una semana, manteniendo el registro de si la venta pertenece a las tiendas en físico o fue realizada en línea, e indicando tambien si en esa semana empezó o no la iniciativa de Schwarz.

Entonces el modelo es

$$\mathbb{E}[y_{ijk}|\alpha, \beta, (\alpha\beta)] = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ijk}, \text{ con } ij \in \{1, 2\}$$

Luego, con ayuda de R hacemos el curado de datos, así como la creación de nuestras variables (ver link en github), y el ajuste del modelo con el comando $lm()$.

Salida de R

```
# Call:
#   lm(formula = sales_per_week ~ afterBops * isSaleOnline)
#
# Residuals:
#   Min       1Q   Median       3Q      Max
# -2105010  -392746   -59945   213084  3332767
#
# Coefficients:
#   Estimate Std. Error
# (Intercept)               1388099     150950
# afterBops1                -214938     215600
# isSaleOnline1             3157482     211490
# afterBops1:isSaleOnline1  -239695     300612
```

```
#  t value  Pr(>|t|)
# (Intercept)                9.196   5.3e-15 ***
#   afterBops1              -0.997      0.321
# isSaleOnline1             14.930   < 2e-16 ***
# afterBops1:isSaleOnline1  -0.797      0.427
# ---
#   Signif. codes:
#   0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 769700 on 101 degrees of freedom
# (1 observation deleted due to missingness)
# Multiple R-squared:  0.8037,Adjusted R-squared:  0.7979
# F-statistic: 137.9 on 3 and 101 DF,  p-value: < 2.2e-16
```

El ajuste del modelo es confiable, podríamos decir, dado que el $R^2 = 0.8037$ lo que nos señala que se explica un 80.37% de la variación total de los datos.

Con base a la salida y abusando un poco de la notación, para la mejor comprensión de los resultados, tenemos que

$$\begin{aligned}
\mathbb{E}[y|\text{antes de Bops, venta en físico}] &= 1,388,099 \\
\mathbb{E}[y|\text{despues de Bops, venta en físico}] &= 1,173,161 \\
\mathbb{E}[y|\text{antes de Bops, venta en línea}] &= 4,545,581 \\
\mathbb{E}[y|\text{despues de Bops, venta en línea}] &= 4,090,948
\end{aligned}$$

Si comparamos los casos del antes y despues de que se lanzara la iniciativa en cada división, vemos claramente que las ventas semanales de la compañía decayeron en un 15.5% en la división de las tiendas en físico mientras que en la división de las ventas en línea decayeron un 10%.

¿Cuántos millones de dólares se ganaron o perdieron a partir del programa? Si hacemos un pequeño calculo, el diferencial del valor esperado en ventas, del antes y después de la iniciativa, fue de $214,938$ dólares para la división de ventas en físico, por otro lado para la división de ventas en línea fue de $454,633$ dólares, lo que nos da una pérdida total de $669,571$ dólares por semana en toda la compañía.

¿Deberían expandirse a Canadá? Mi respuesta sería no, dado los análisis ya expuestos.

Github

El codigo fuente de cada uno de los ejercicios se ecnuentra en el siguiente link:

`https://github.com/Eduardosh9324/opi_analytics`

Los ejercicios fueron realizados tanto en R como en Python, (por la comodidad de resolver el problema en cierto lenguaje).Como comentario adicional, tuve problemas de mostrar el código fuente de la parte de python, por lo que está tanto un pdf de la página donde se depliega el archivo *.ipynb* como este mismo.