ABRAXAS

Test

Carlos Alberto Juárez Santini
jusca_94@hotmail.com

Contenido

Objetivo de la prueba

In this competition, you will forecast the demand of a product for a given week, at a particular store. The dataset you are given consists of 8 weeks of sales transactions in Mexico.

Every week, there are delivery trucks that deliver products to the vendors. Each transaction consists of sales and returns. Returns are the products that are unsold and expired. The demand for a product in a certain week is defined as the sales this week subtracted by the return next week.

Limpieza de datos

Buscamos si hay algún tipo de Missing Value, Null o NaN. En caso de existir algún dato de estos, eliminamos esa instancia.

Analysis Exploratory

Tenemos 4 archivos, los cuales utilizaremos como datasets o base de datos:

- df_[candidate]_small.csv
- df_[test]_small.csv
- producto_tabla.csv
- town_state_small.csv

Obtenemos una descripción de los datos que componen cada uno de estos archivos. De igual se obtiene el nombre de las columnas de cada uno de estos datasets.

- df_[candidate]_small

```
RangeIndex: 7974418 entries, 0 to 7974417
Data columns (total 11 columns):
 #   Column             Dtype
---  ------             -----
 0   Semana             int64
 1   Agencia_ID         int64
 2   Canal_ID           int64
 3   Ruta_SAK           int64
 4   Cliente_ID         int64
 5   Producto_ID        int64
 6   Venta_uni_hoy      int64
 7   Venta_hoy          float64
 8   Dev_uni_proxima    int64
 9   Dev_proxima        float64
 10  Demanda_uni_equil  int64
```

	Semana	Agencia_ID	Canal_ID	Ruta_SAK	Cliente_ID	Producto_ID	Venta_uni_hoy	Venta_hoy	Dev_uni_proxima	Dev_proxima	Demanda_uni_equil
0	3	1110	7	3301	15766	1212	3	25.14	0	0.0	3
1	3	1110	7	3301	15766	1216	4	33.52	0	0.0	4
2	3	1110	7	3301	15766	1238	4	39.32	0	0.0	4
3	3	1110	7	3301	15766	1240	4	33.52	0	0.0	4
4	3	1110	7	3301	15766	1242	3	22.92	0	0.0	3

- df_[test]_small

```
RangeIndex: 1337913 entries, 0 to 1337912
Data columns (total 6 columns):
 #   Column       Non-Null Count    Dtype
---  ------       --------------    -----
 0   Semana       1337913 non-null  int64
 1   Agencia_ID   1337913 non-null  int64
 2   Canal_ID     1337913 non-null  int64
 3   Ruta_SAK     1337913 non-null  int64
 4   Cliente_ID   1337913 non-null  int64
 5   Producto_ID  1337913 non-null  int64
```

	Semana	Agencia_ID	Canal_ID	Ruta_SAK	Cliente_ID	Producto_ID
0	9	1110	7	3301	15766	1212
1	9	1110	7	3301	15766	1238
2	9	1110	7	3301	15766	1240
3	9	1110	7	3301	15766	1242
4	9	1110	7	3301	15766	1250

- producto_tabla

```
RangeIndex: 2592 entries, 0 to 2591
Data columns (total 2 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Producto_ID     2592 non-null   int64
 1   NombreProducto  2592 non-null   object
```

	Producto_ID	NombreProducto
0	0	NO IDENTIFICADO 0
1	9	Capuccino Moka 750g NES 9
2	41	Bimbollos Ext sAjonjoli 6p 480g BIM 41
3	53	Burritos Sincro 170g CU LON 53
4	72	Div Tira Mini Doradita 4p 45g TR 72

- town_state_small

```
RangeIndex: 41 entries, 0 to 40
Data columns (total 3 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Agencia_ID  41 non-null     int64
 1   Town        41 non-null     object
 2   State       41 non-null     object
```

	Agencia_ID	Town	State
0	1110	2008 AG. LAGO FILT	MÉXICO, D.F.
1	1111	2002 AG. AZCAPOTZALCO	MÉXICO, D.F.
2	1112	2004 AG. CUAUTITLAN	ESTADO DE MÉXICO
3	1113	2008 AG. LAGO FILT	MÉXICO, D.F.
4	1114	2029 AG.IZTAPALAPA 2	MÉXICO, D.F.

Correlacion entre variables

Se hizo una correlación (ver Figura 1) entre las variables del dataset *df_[candidate]_small*, en el cual tenemos como datos categóricos las variables Semana, Agencia_ID, Canal_ID, RutaSAK, Cliente_ID y Producto_ID. Estas variables lo ideal es seguirlas manejando en este formato para más adelante poder usarlas con mayor facilidad en la exploración de los datos e incluso en la creación del modelo de forecast. El resto de variables son numéricas.

Se observa que existe una fuerte relación positiva entre las ventas las cuales están altamente correlacionadas con la demanda. Dado que la demanda se obtiene a partir de ventas menos devoluciones, es decir $demanda = ventas - devoluciones$.

Figura 1. Correlación entre las variables del dataset df_[candidate]_small.

En la Tabla 1 se observa que la Demanda (Demanda_uni_equil) y las Ventas (Venta_uni_hoy) son básicamente lo mismo. Podemos discutir en que su diferencia puede estar dada a la existencia de devoluciones. Aunque si vemos las devoluciones, estas prácticamente son cero.

Esto nos puede dar indicio de que muy pocas veces hay devoluciones por parte de los clientes. Este factor puede llegar a complicar el forecast de la demanda. Es decir que la demanda del cliente por un producto, esta puede estar localizada entre 0 y 4 productos, pero puede llegar a 4700.

Vemos que hay una gran variación en los datos reales de la demanda y su desviación estándar se localiza en 22 aproximadamente. Como se tienen muy pocas devoluciones, esto puede complicarnos para el forecast ya que también es un dato importante para calcular la demanda.

	Venta_uni_hoy	Venta_hoy	Dev_uni_proxima	Dev_proxima	Demanda_uni_equil
Media	7.534587	81.407380	0.104409	1.209538	7.460012
Mediana	4.000000	34.590000	0.000000	0.000000	4.000000
Minimo	0.000000	0.000000	0.000000	0.000000	0.000000
Maximo	4800.000000	647360.000000	3360.000000	49500.000000	4732.000000
Moda	2.000000	16.760000	0.000000	0.000000	2.000000
Varianza	486.781687	343354.322487	10.315230	802.281446	479.937775
Std	22.063130	585.964438	3.211733	28.324573	21.907482

Tabla 1. Descripción estadística de las variables Venta_uni_hoy, Venta_hoy, Dev_uni_proxima, Dev_proxima y Demanda_uni_equil.

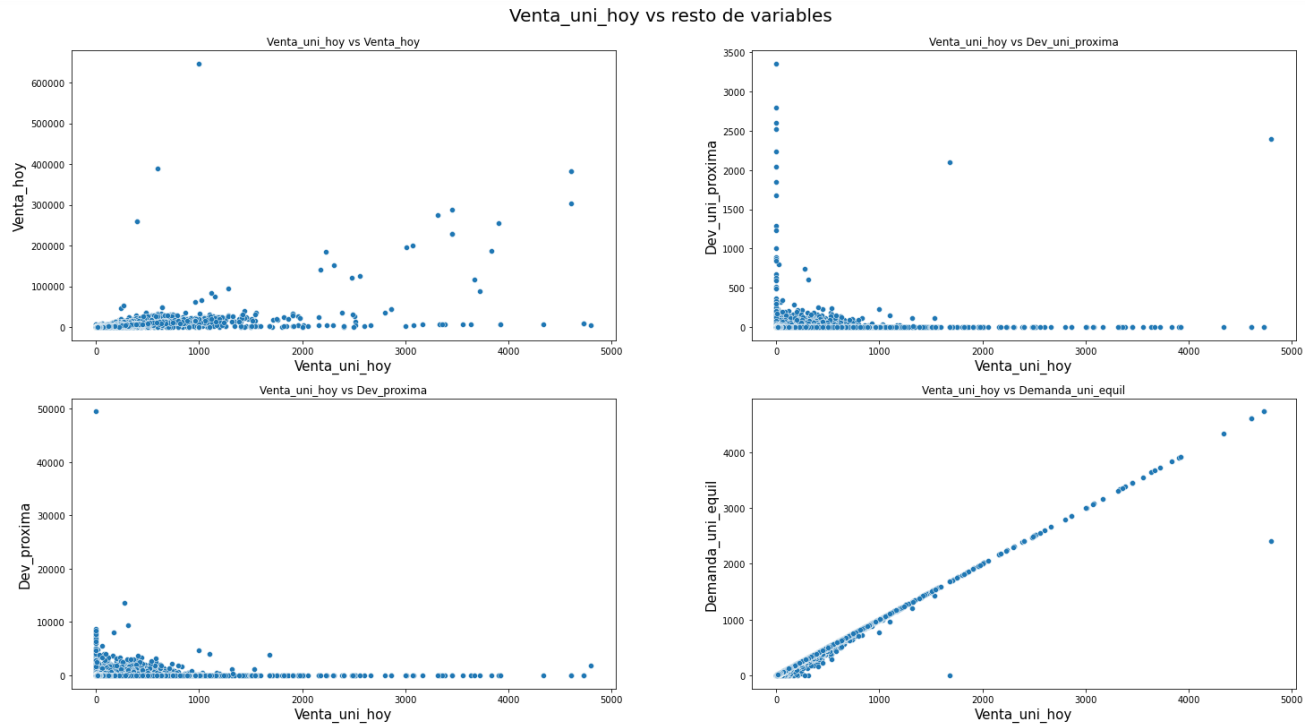

Figura 2. Grafica de dispersión entre la variable Venta_uni_hoy comparada con el resto de variables.

Podemos eliminar más adelante la variable de Venta_uni_hoy que está fuertemente correlacionada a la demanda. Esto se observa tanto en la Figura 1 y 2. Además vemos (en Figura 2) que la variable Venta_uni_hoy no tiene suficiente correlación con el resto de las variables.

Relación de las variables con la Demana por semana

A continuación, se presenta un análisis de cómo afectan las principales variables categóricas con respecto a la demanda entre el total de semanas y su comportamiento a través de cada semana.

Ruta_SAK vs Demanda_uni_equil

En la Figura 3.1 observamos como hay una gran densidad de demandas en las Ruta_SAK que ronda aproximadamente entre 1 y 500. Luego disminuye y vuelve a crecer un tanto entre las Ruta_SAK 10 a 2100. Aproximadamente entre la Ruta_SAK 3000 y 3500 vuelve a centrarse un buen volumen de demandas, no tanto como en la 1 a 500, pero se observa que la demanda crece unas pocas ocasiones arriba de 3000. En cuanto a las Ruta_SAK localizadas cerca de 10,000 las demandas de estas rutas son muy bajas en comparación con las rutas menores a 8000.

Si bien este comportamiento en general durante las semanas 3 a 8 no nos brinda un entendimiento especifico de cómo afecta la demanda utilizando solamente la variable de Ruta_SAK.

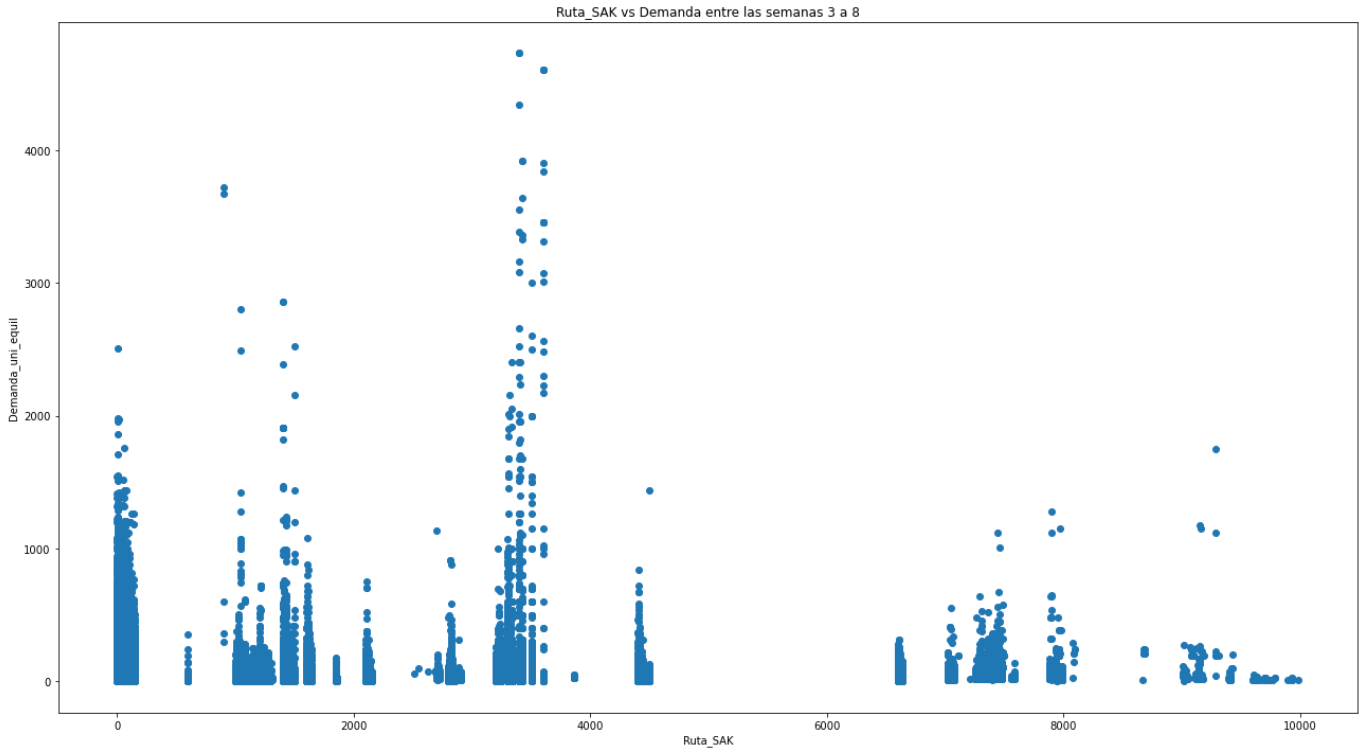

Figura 3.1. Distribución de las Ruta_SAK con respecto a la Demanda a lo largo de las semanas 3 a 8.

En la Figura 3.2 se observa una gráfica de distribución de las Ruta_SAK contrastada a la Demanda para cada una de las semanas.

En todas las semanas se detecta como patrón característico que para las Ruta_SAK entre 1 y 500 hay una basta densidad sobre la demanda. En general se repite el mismo comportamiento encontrada en todas las semanas para cada una de las 6 semanas.

En la semana 7 se visualiza una disminución de las demandas para todas las rutas en contraste al resto de semanas.

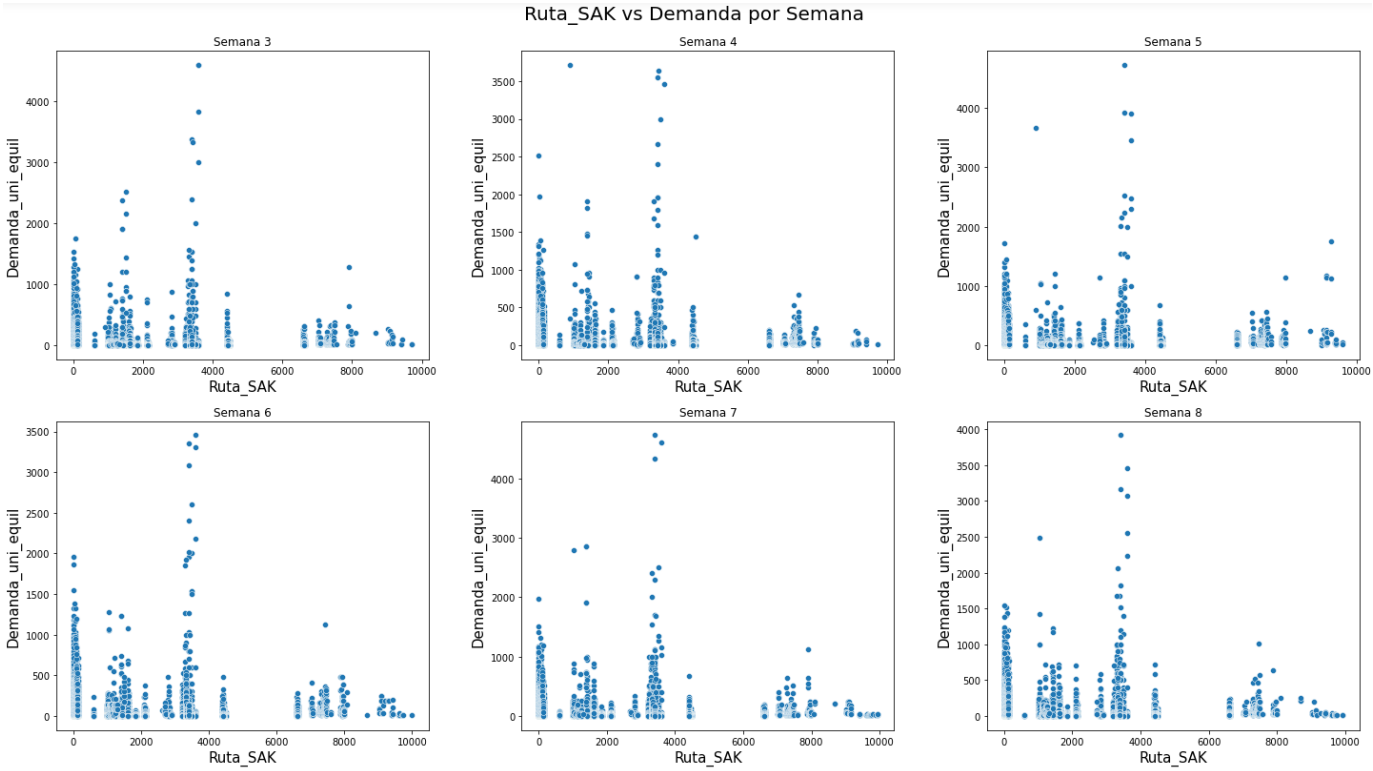

Figura 3.2. Distribución de las Ruta_SAK con respecto a la Demanda en cada semana.

Cliente_ID vs Demanda_uni_equil

En la Figura 3.3 se observa cómo está la distribución de los Cliente_ID con respecto a la Demanda, vemos casi todos los Cliente_ID se localizan en 1 a 100000, solo hay unos cuentos cercanos al Cliente_ID 2000000 y 10000000 pero estos tienen muy poca demanda de productos.

Podemos ver que la mayoría de la gente tiende a comprar productos en cantidades iguales. Por lo tanto, la naturaleza posiblemente sesgada de la curva de demanda sigue siendo la misma.

Este comportamiento se observa más a detalle a lo largo de cada semana, donde los Cliente_ID más cercanos al origen son los que tienen mayor consumo de productos. Durante las semanas 3,4,5 el cliente que estaba al extremo derecho del eje x ha dejado de consumir productos en las siguientes semanas.

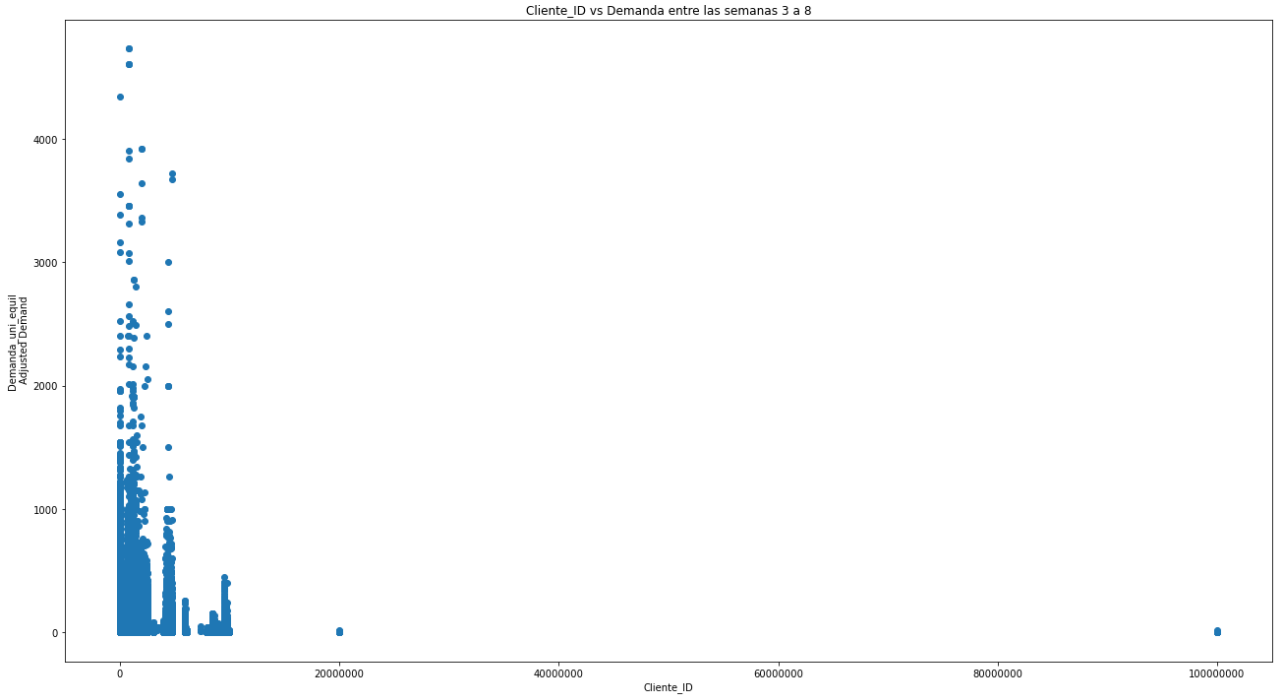

Figura 3.3. Distribución de los Cliente_ID con respecto a la Demanda a lo largo de las semanas 3 a 8.

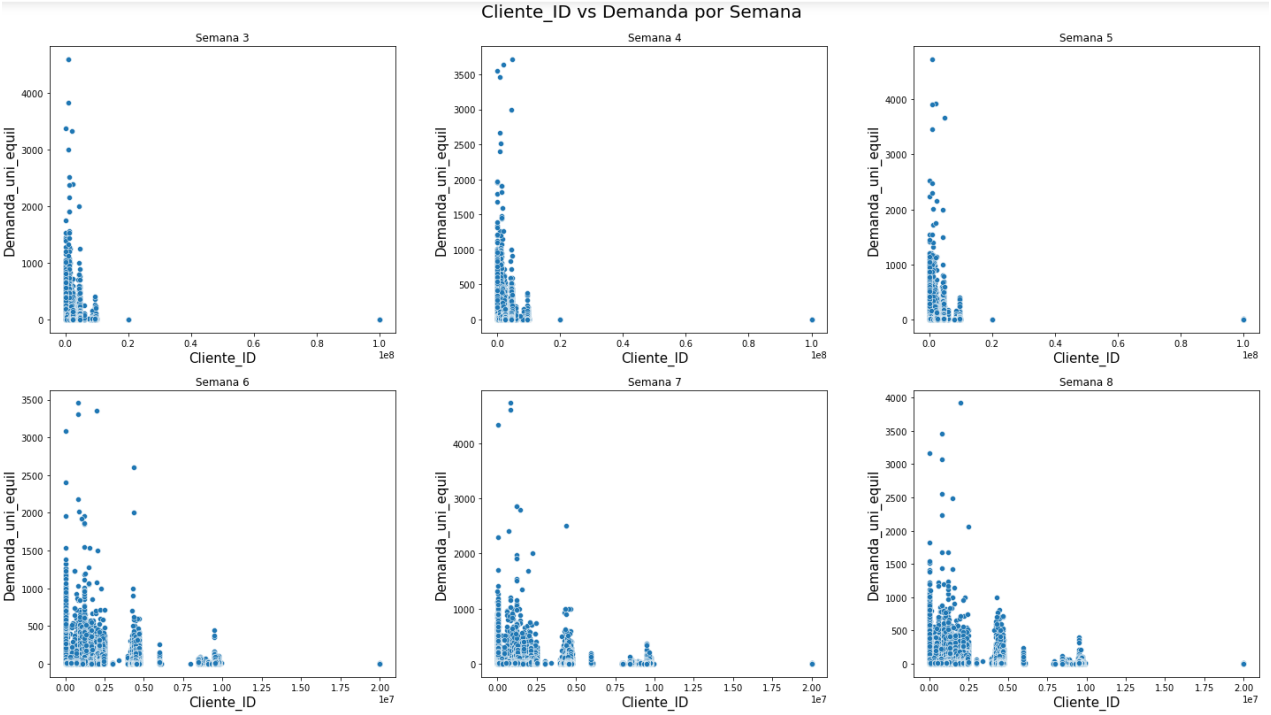

Figura 3.4. Distribución de los Cliente_ID con respecto a la Demanda en cada semana.

Producto_ID vs Demanda_uni_equil

En la figura 3.5 se muestra cómo se distribuyen los productos conforme a la demanda que se tienen. Los Productos_ID de 30000 a 38000 aproximadamente son los que tuvieron mayor demanda, se visualiza una mayor densidad en ese rango, caso contrario sucede entre 0 a 10000 y lo mismo 40000 a 50000 incluso en estos se ven discontinuos los puntos con respecto al eje y.

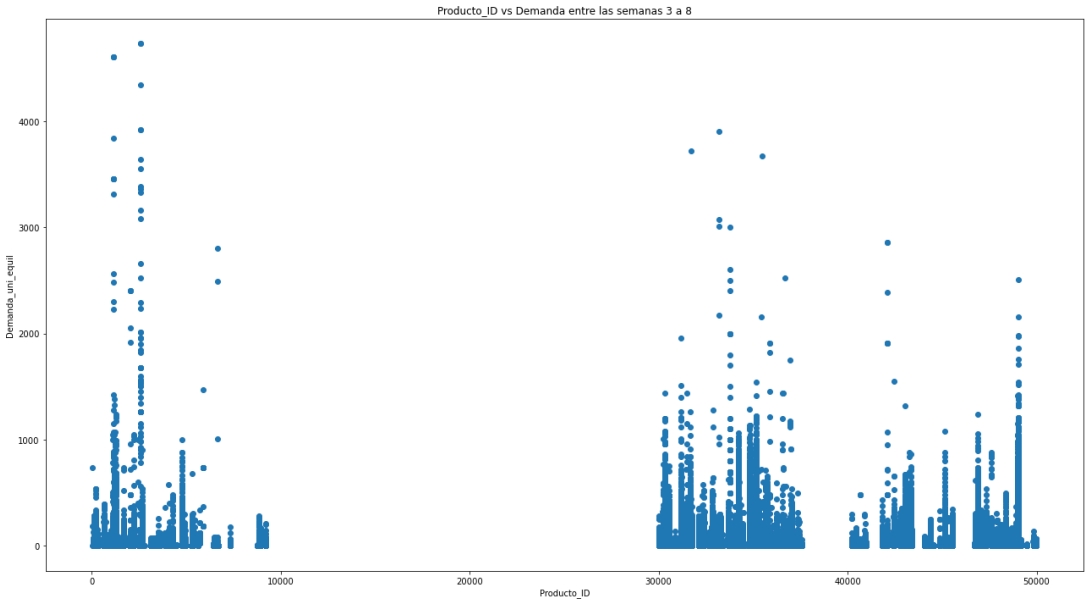

Figura 3.5. Distribución de Producto_ID con respecto a la Demanda a lo largo de las semanas 3 a 8.

En la Figura 3.6 observamos que el comportamiento es muy similar durante cada una de las semanas en los Producto_ID arriba de 30000. A la semana 5 y 7 se nota una disminución entre los Productos_ID de 1 a 1000 en comparación con las demás semanas.

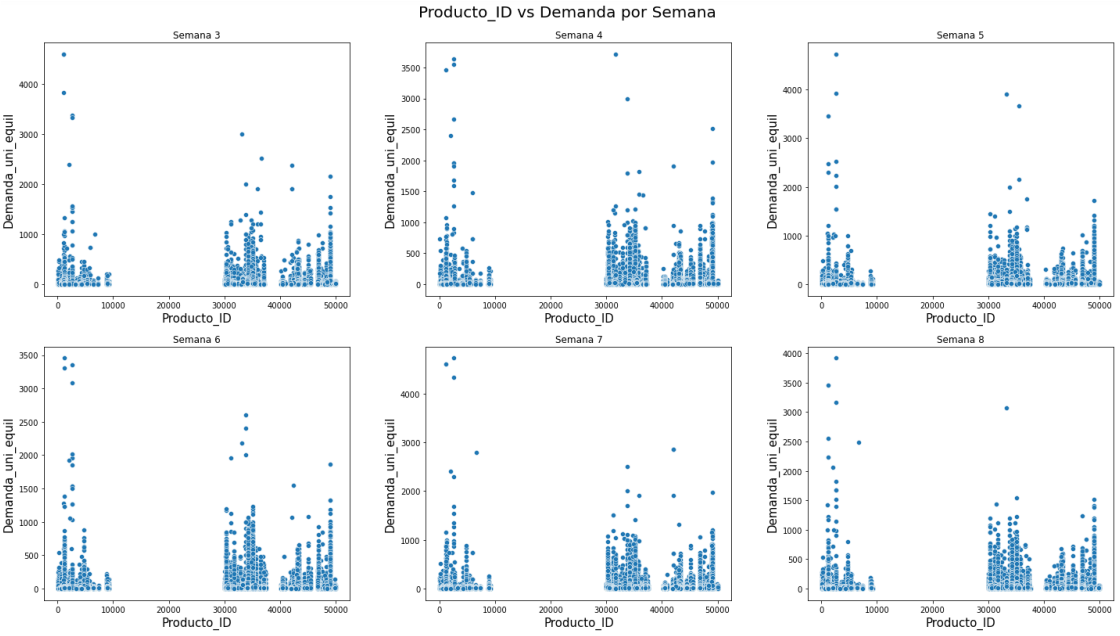

Figura 3.6. Distribución de los Producto_ID con respecto a la Demanda en cada semana.

Canal_ID vs Demanda_uni_equil

En la Figura 3.7 se muestra la relación que existe entre Canal_ID y Demanda_uni_equil entre las semanas 3 a 8. Se visualizan 7 Canales, de los cuales el canal 1 y 2 son los que registran la mayoría de las ventas.

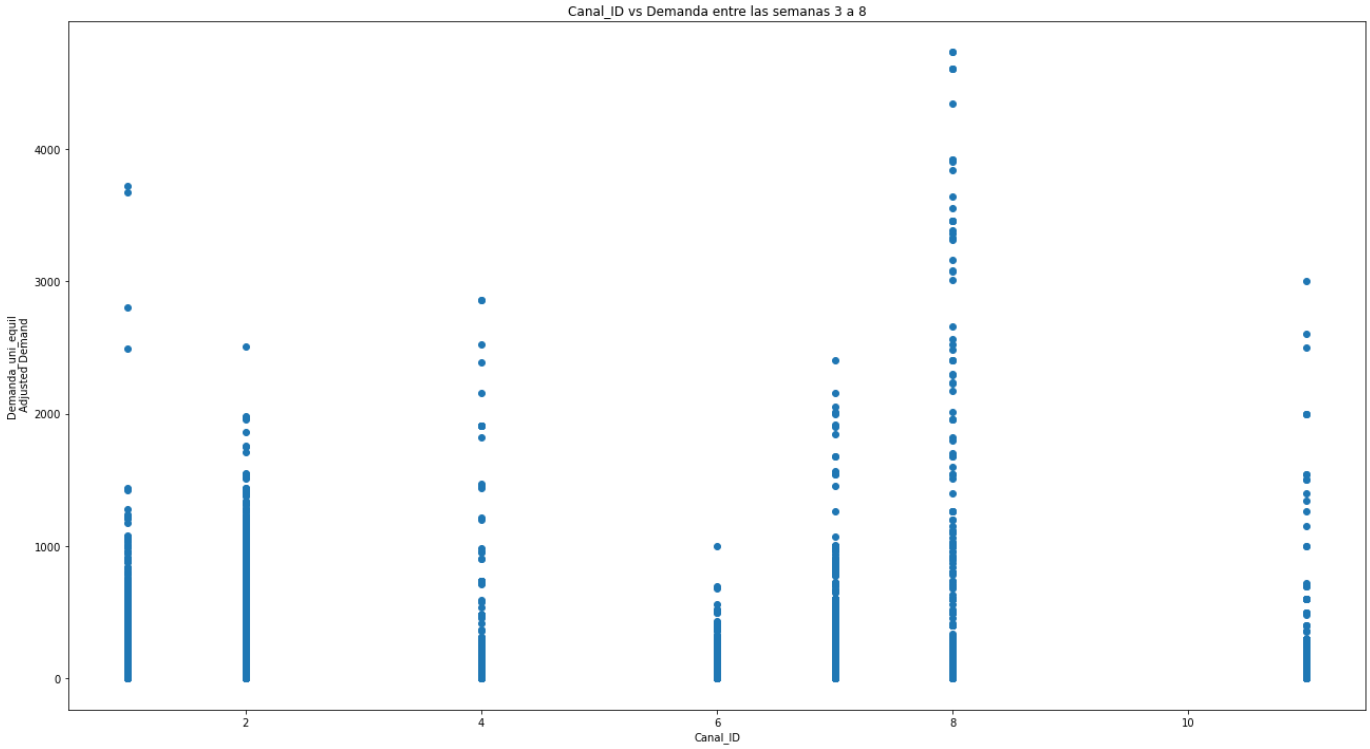

Figura 3.7. Distribución de Canal_ID con respecto a la Demanda a lo largo de las semanas 3 a 8.

Este comportamiento se observa en la Figura 3.8 el mismo patrón durante cada semana, el cual hacemos referencia para los Canal_ID 1 y 2. Son los que registran mas ventas. Caso contrario con el Canal 4 que en las semanas 5 y 8 sus ventas son bajas, esto no se visualiza en la gráfica general, pero si en las que muestra resultados semanales.

Con esto vemos que la distribución de los datos es muy baja, por lo que no podemos llegar a una conclusión sobre la demanda dependiendo únicamente del Canal_ID.

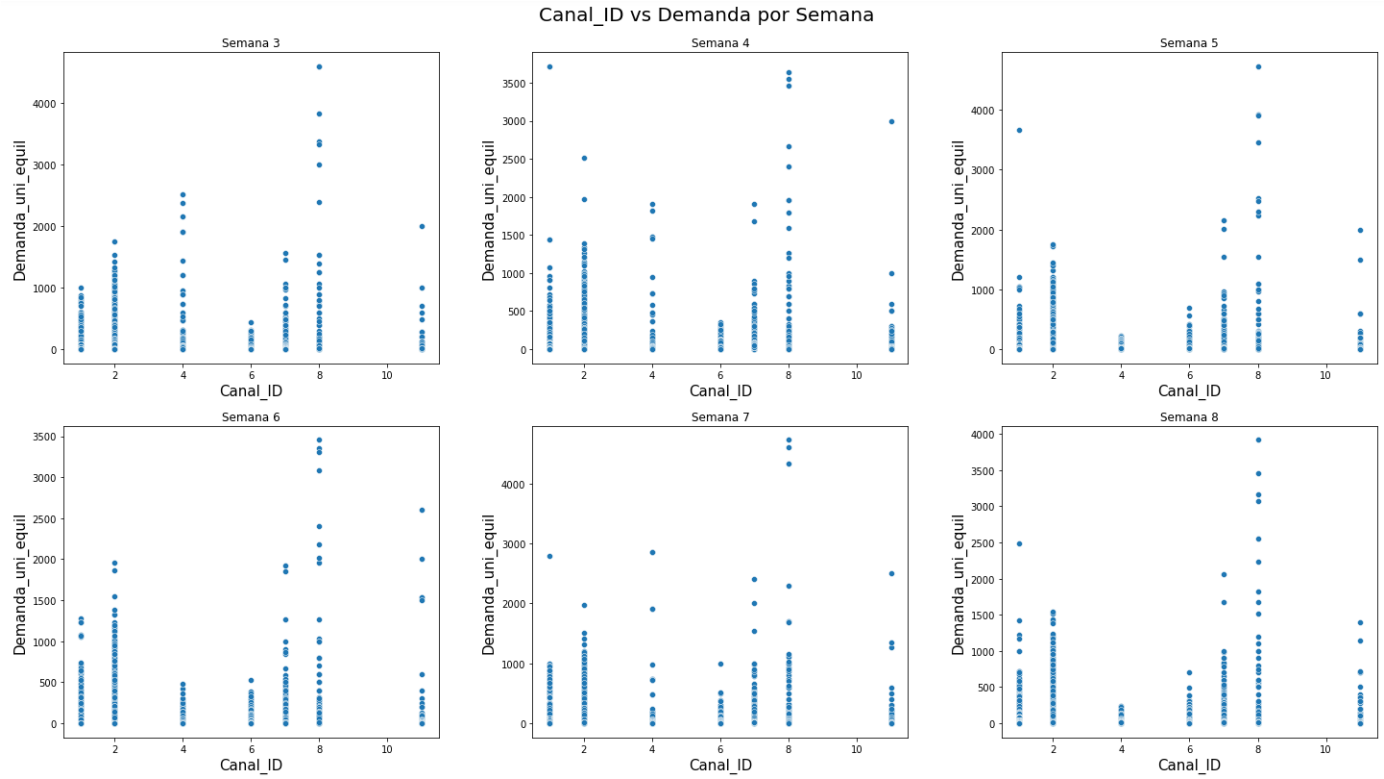

Figura 3.8. Distribución de los Canal_ID con respecto a la Demanda en cada semana.

Agencia_ID vs Demanda_uni_equil

La Figura 3.9 nos detalla la relación existente entre Agencia_ID y la Demanda a lo largo de las semanas. En este vemos que en las primeras agencias la distribución es mayor comparada a las que rodean a la agencia 1120. Después se observa una distribución discontinua en las agencias alrededor de la 1150.

La Figura 3.10 muestra la relación de estas dos variables, pero para cada semana. En el cual el comportamiento de los datos es muy similar en cada semana comparado al registro total de todas las semanas. Incluso se observa que en las agencias 1120 y sus cercanas a esta tienen mayor densidad en la distribución de los datos.

Con esto, la Agencia_ID no ayuda por si sola a hacer un forecast para la demanda.

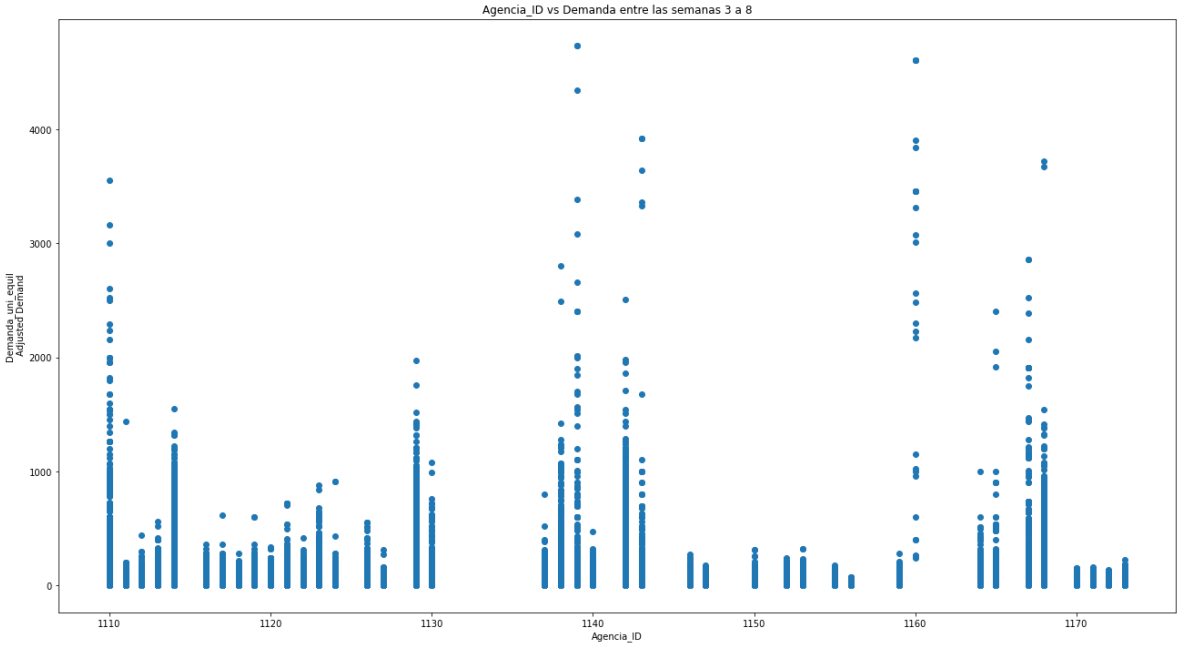

Figura 3.9. Distribución de Agencia_ID con respecto a la Demanda a lo largo de las semanas 3 a 8.

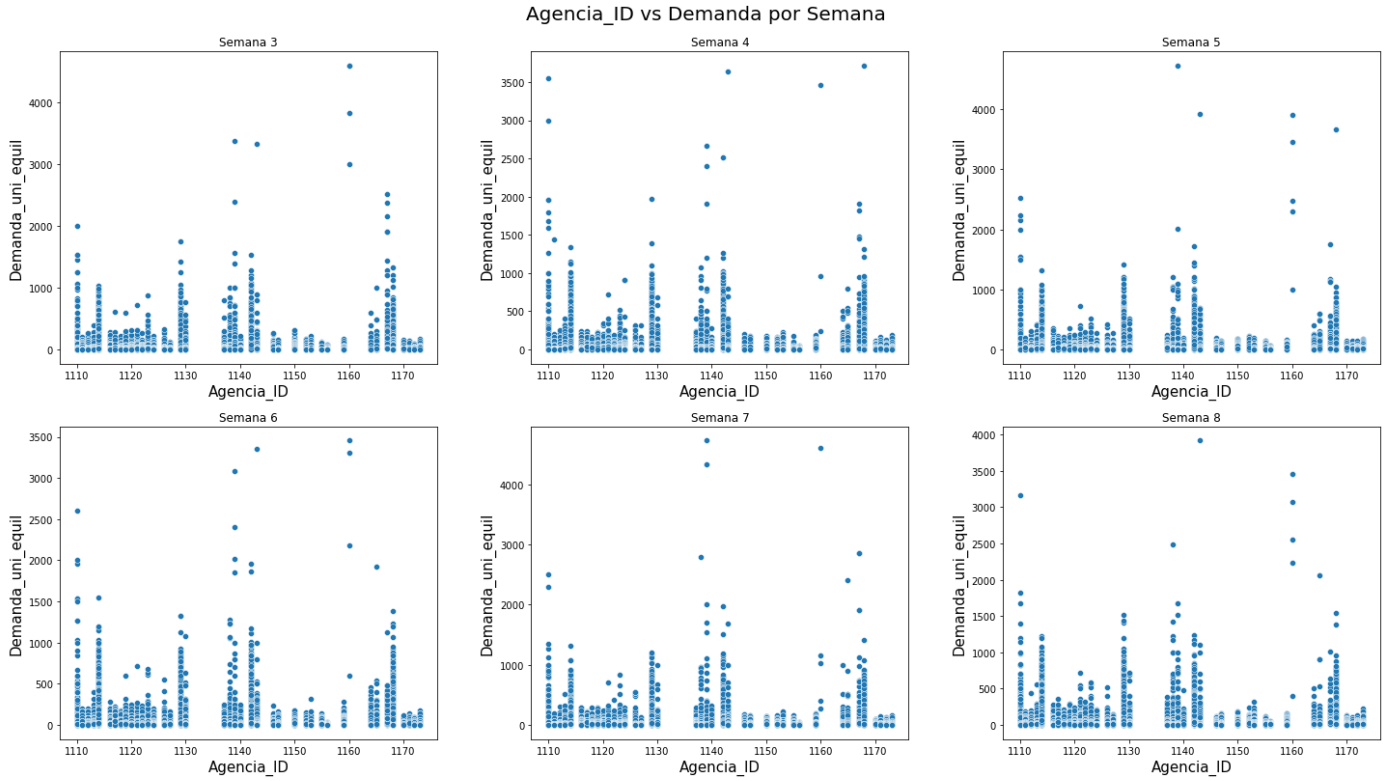

Figura 3.10. Distribución de los Canal_ID con respecto a la Demanda en cada semana.

Conclusión de uso de los atributos para hacer la predicción de la demanda

De acuerdo con las descripciones detalladas en los puntos anteriores, podemos ver claramente que los atributos individuales por sí solos no nos ayudan a predecir la demanda. Por lo tanto, consideramos una combinación de atributos.

Una de estas combinaciones puede ser detectar Cliente_ID y Producto_ID para cada semana. Dado que los clientes tienden a consumir los mismos productos una y otra vez. Esta consideración puede resultar muy buena siempre y cuando no ingresen nuevos productos a los registros, puesto que no tenemos registro de como los clientes puedan actuar con la incorporación de los nuevos productos.