# Proyecto Integrador III

## Crea un notebook, carga los csv. responde y justifica.

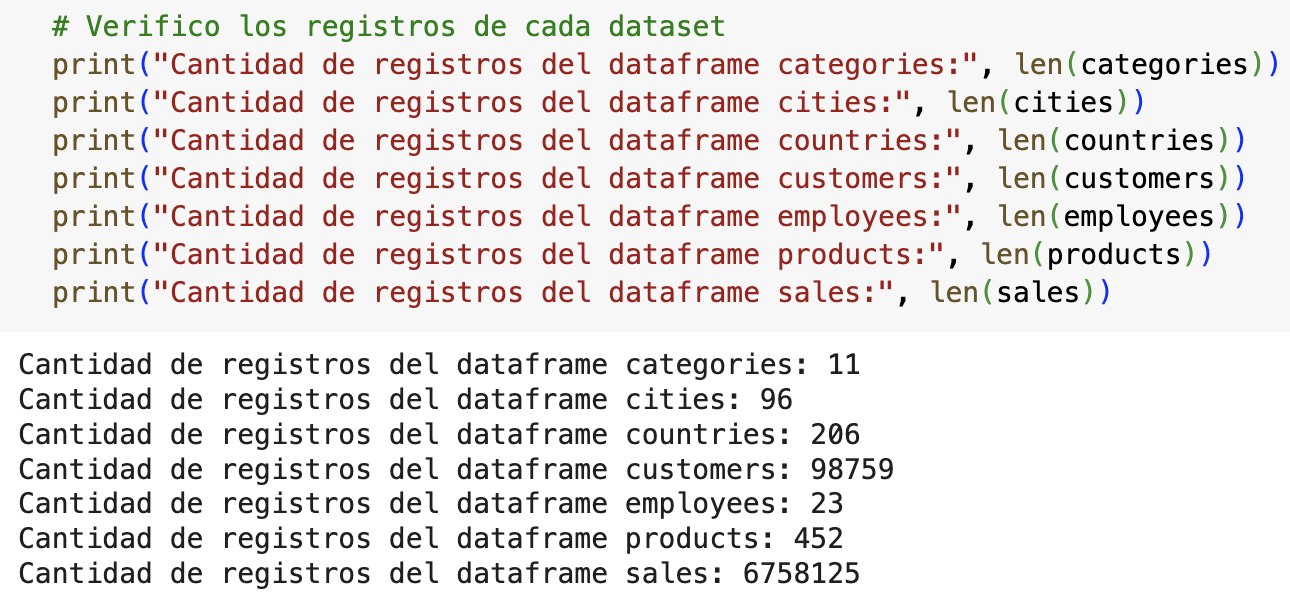
El notebook se debe entregar y debe ser posible visualizar los resultados de cada pregunta:

El campo TotalPrice en la tabla sales no tiene valores válidos. Utilizando la información de precios de la tabla products, calcula el valor real de la venta para cada registro y almacena en una nueva columna

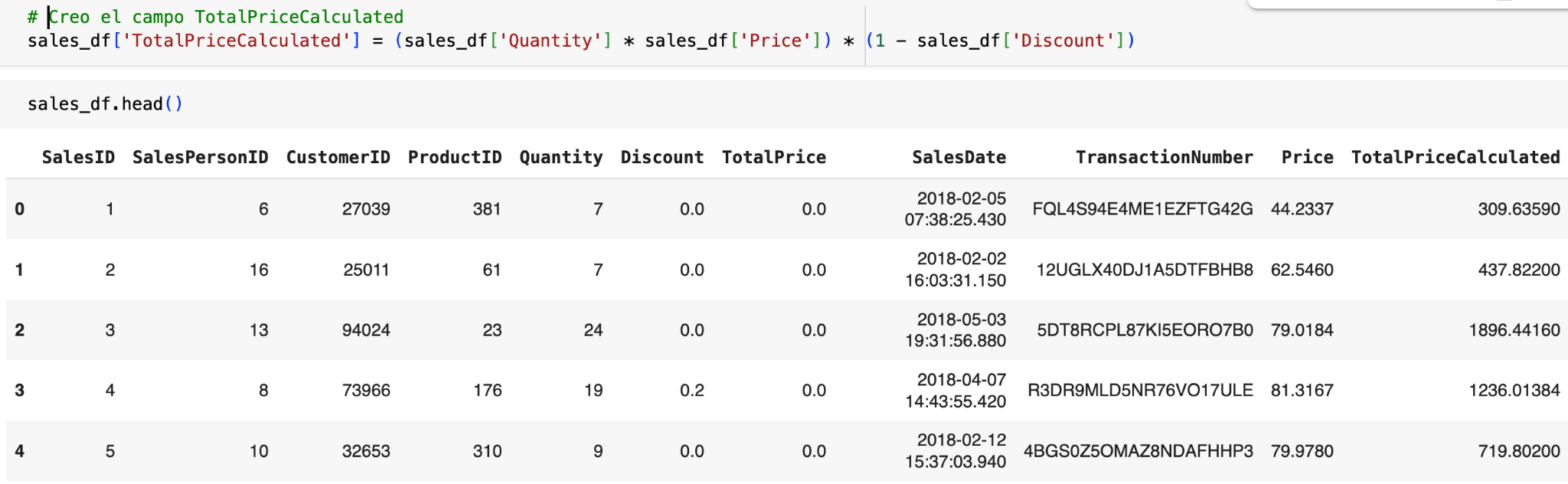
Utiliza la siguiente fórmula:

TotalPriceCalculated=(Quantity×UnitPrice)×(1−Discount)

**Evidencia:**

****

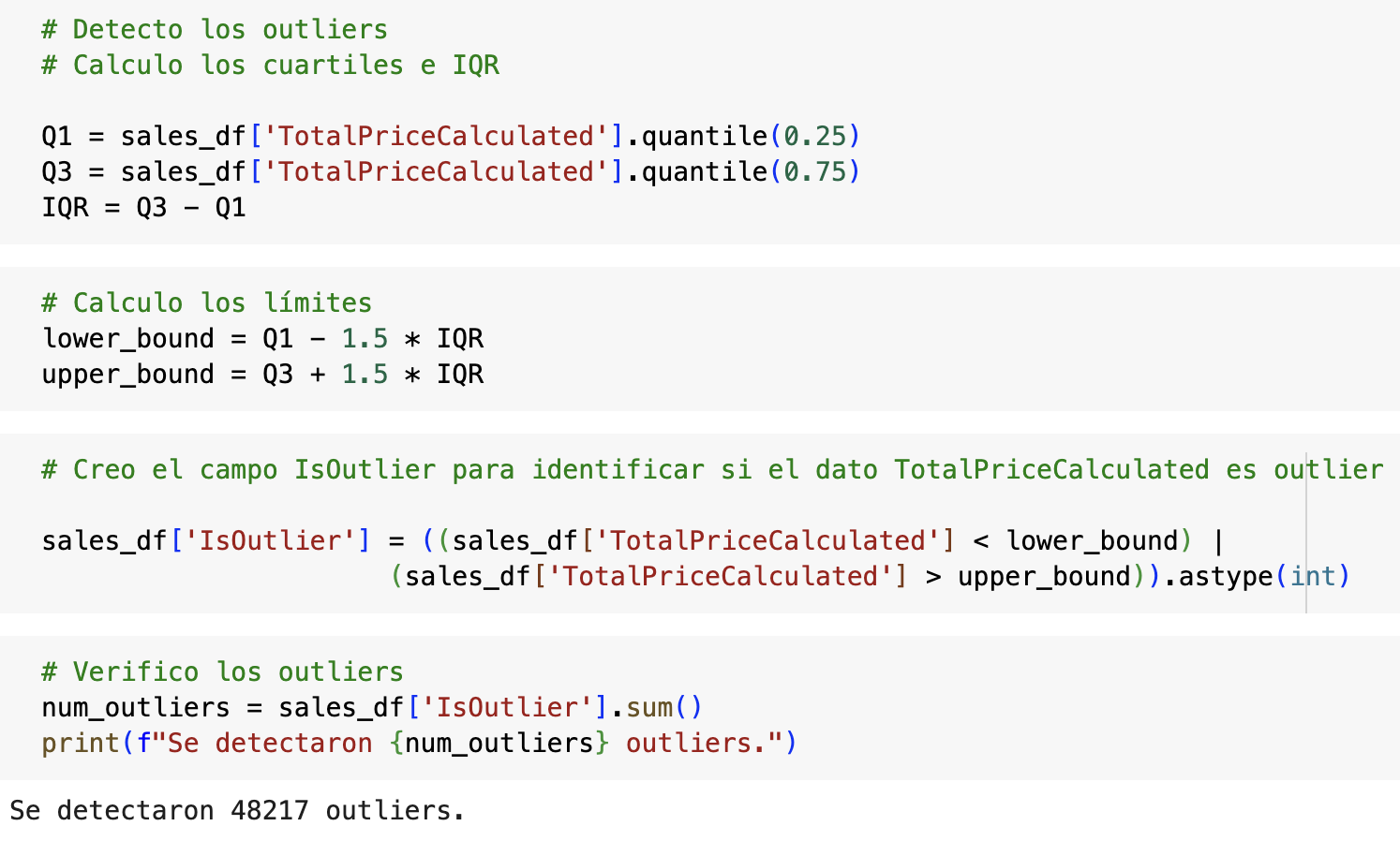
Acá se puede ver la carga de los diferentes dataframes desde los archivos indicados.



En esta imagen se puede ver el nuevo campo calculado llamado **TotalPriceCalculated**

## Detecta los outliers en la columna de ventas totales (TotalPriceCalculated)

Utilizando el criterio del rango intercuartílico (IQR). Luego, crea una nueva columna llamada IsOutlier que tenga el valor 1 si el registro es un outlier y 0 en caso contrario. ¿Cuántos outliers se detectaron?



Según el dataframe suministrado se han detectado 48217 datos outliers.

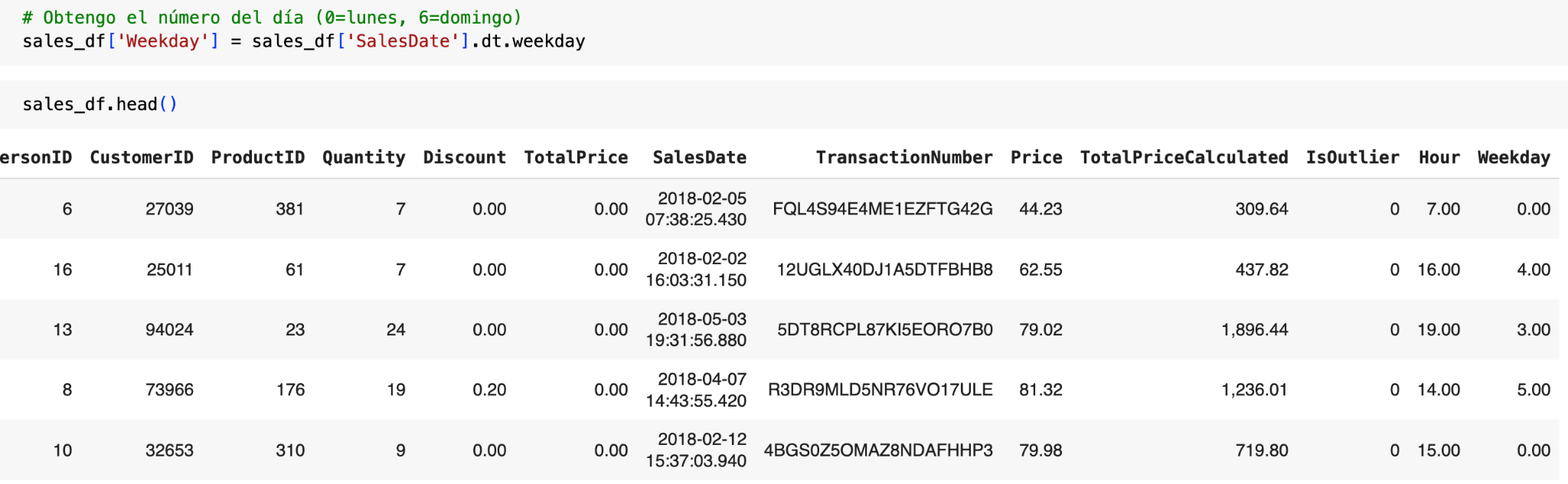
## A partir de la columna SalesDate, crea una nueva columna que contenga únicamente la hora de la venta.

Luego, identifica en qué hora del día se concentran más ventas totales (TotalPriceCalculated).

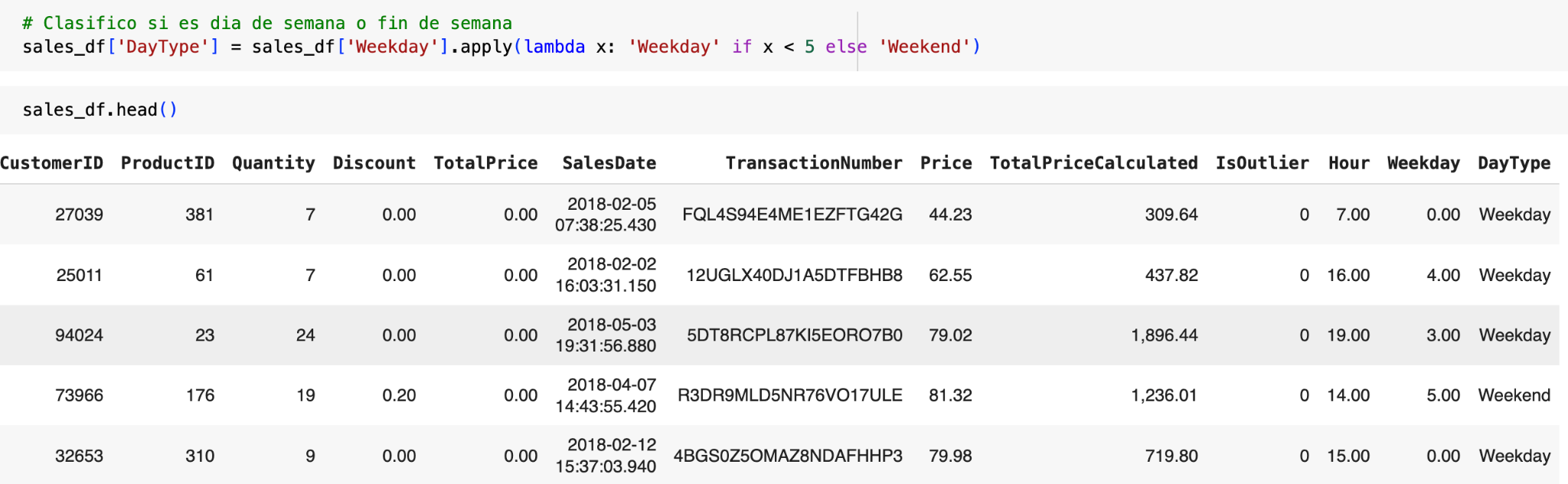


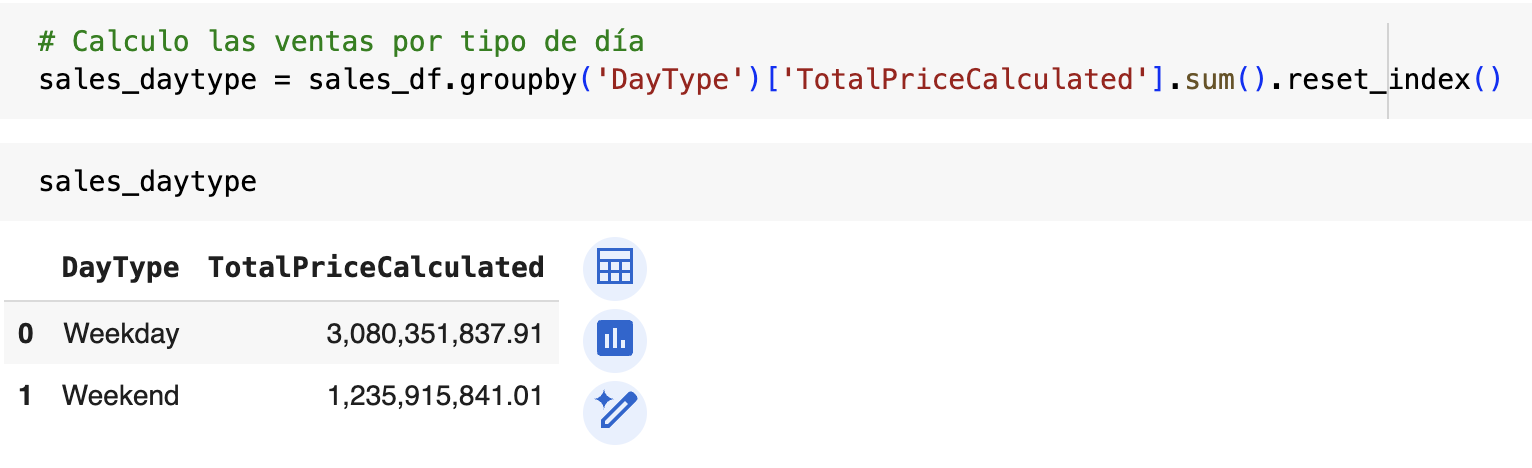
En esta imagen se puede ver la cantidad de ventas por hora. La hora que más se ve es a las 16 horas con una suma total de 179,014,421.24.

¿La empresa vende más durante los días de semana o en el fin de semana? Utiliza la columna SalesDate para identificar el día de la semana de cada venta, clasifica los registros como Entre semana o Fin de semana, y compara el total de ventas (TotalPriceCalculated) entre ambos grupos.



Aquí creo la columna Weekday del dataframe el número del día donde 0 es lunes y 6 es domingo. Una vez calculado el número del día calculo si es dia de la semana o fin de semana como se puede ver en la siguiente imagen.



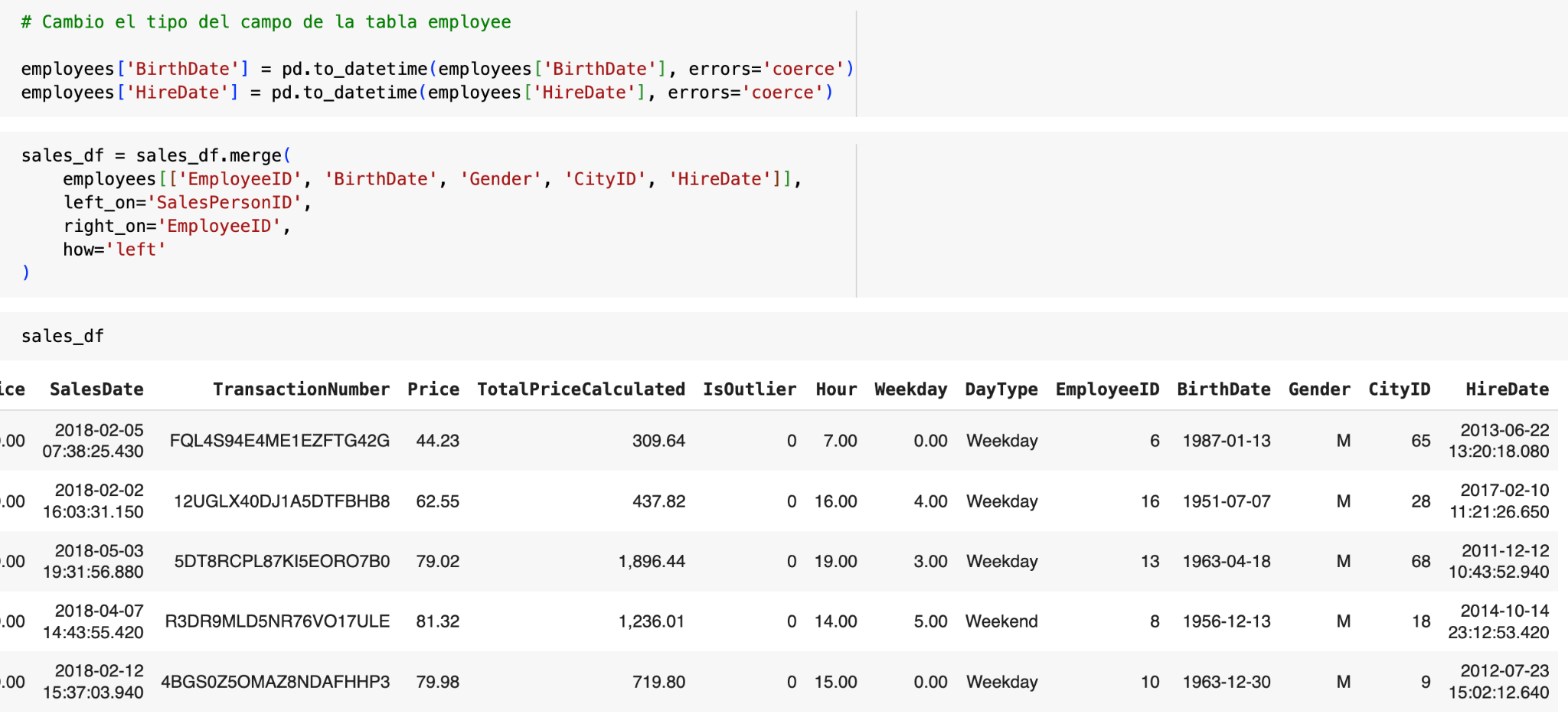


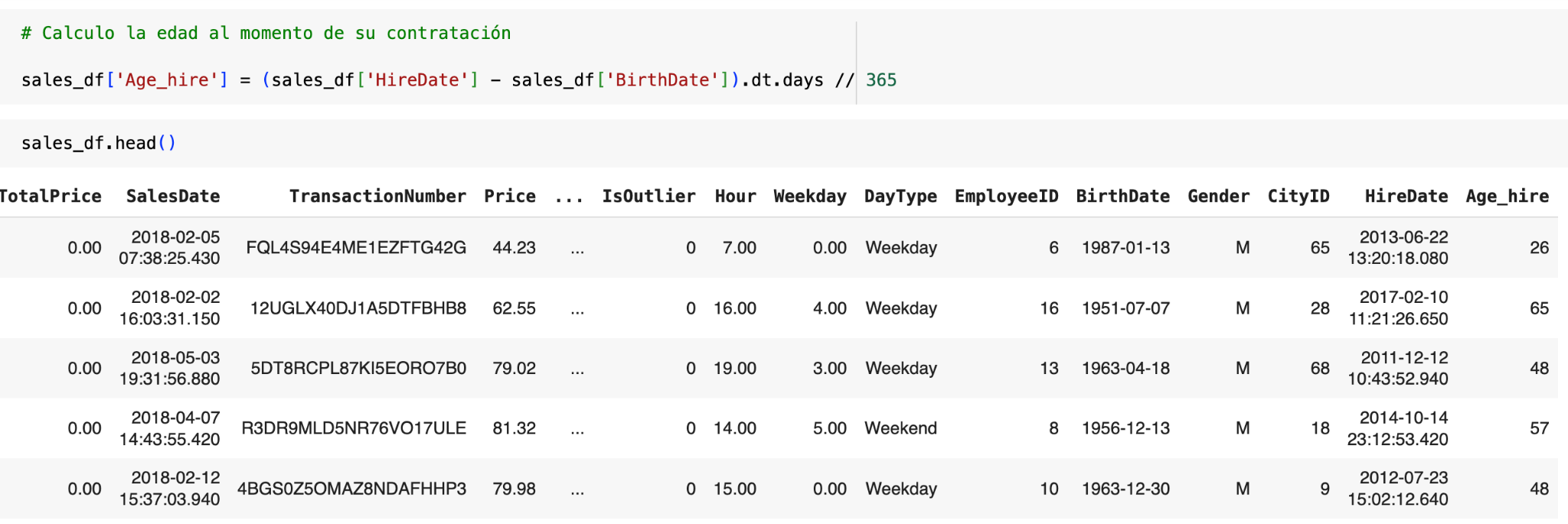
En esta imagen se puede verificar que las mayores compras se realizan durante los días de la semana. El importe total es de 3,080,351,837.91.

## Como parte del proceso de feature engineering, en el mismo df que vienes trabajando, calcula dos nuevas columnas en el dataset de ventas:

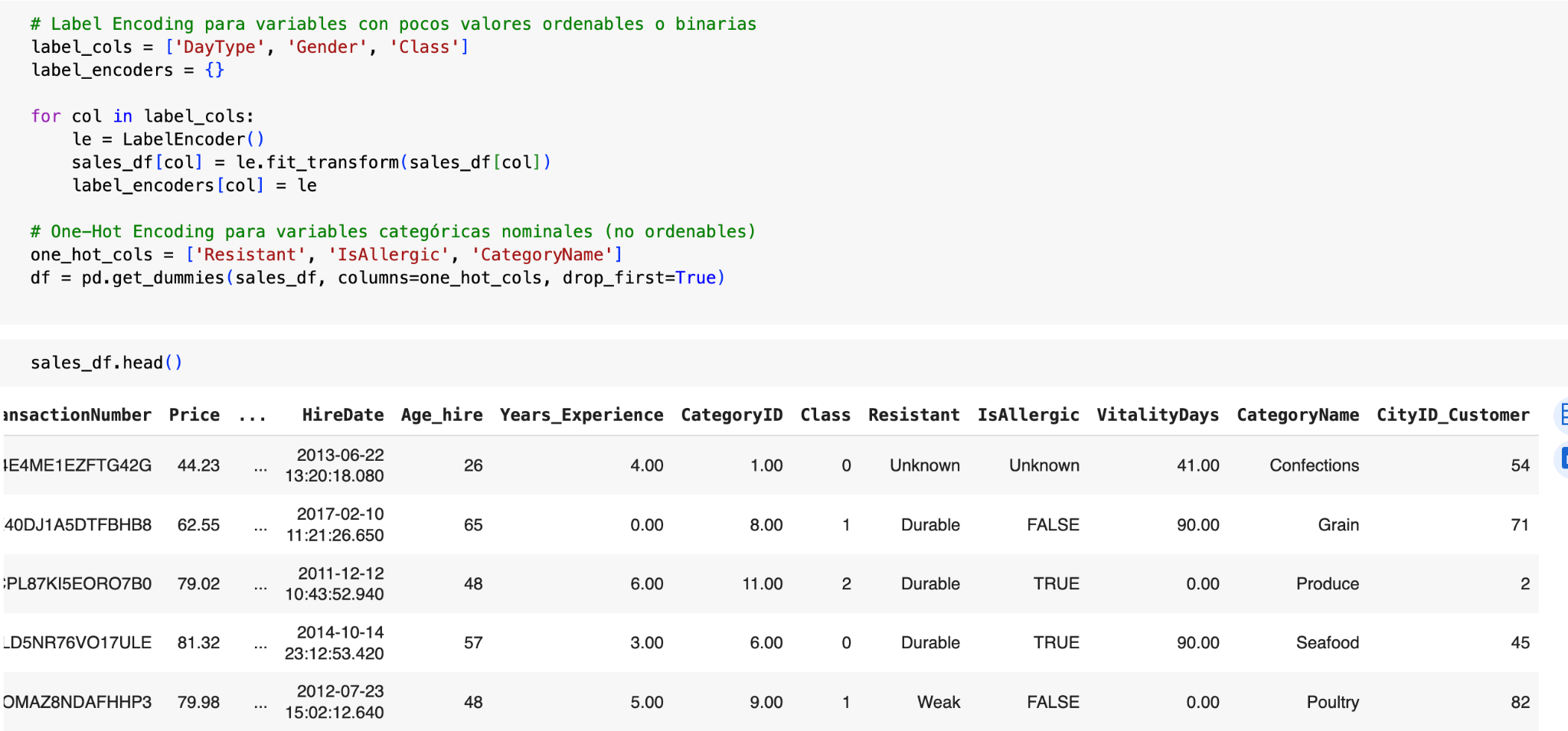
La edad del empleado al momento de su contratación y años de experiencia al momento de realizar cada venta.

Utiliza las columnas BirthDate, HireDate (de la tabla employees) y SalesDate (de la tabla sales). Asegúrate de trabajar con fechas en formato adecuado.









IMPORTANTE

* Prepara un único dataset definitivo para modelado que combine información relevante de las tablas disponibles.
* Incluye las features que se han calculado previamente.
* Aplica transformaciones adecuadas a las variables categóricas y a las variables numéricas (si lo consideras necesario) para dejar los datos listos para ser utilizados por un modelo de machine learning.
* Justifica las transformaciones realizadas. La variable objetivo es TotalPriceCalculated, por lo que debe quedar sin transformaciones.

### Conocimientos necesarios

* Manipulación de datos
* Limpieza
* Transformación
* Feature engineering
* Clasificación temporal

### Tech Stack necesario

* Python
* Pandas
* Jupyter Notebook

# Respuestas