PROYECTO FINAL

Diaz Gastón Alejandro - DNI: 32016726

Dejo este link donde se puede ver y/o descargar los archivos

Dejo este link donde se puede ver un Dashboard Dinamico del EDA

La empresa Sanz Supermercado, está radicado en un barrio de la ciudad de Rio Cuarto, Provincia de Córdoba, en argentina. Esta empresa desea mejorar su logística de los productos que comercializa y aumentar la cantidad de sus ventas, es un supermercado pequeño y la superficie de su local es pequeña.

Dicha empresa tiene varios datos de las compras realizadas por sus clientes, posee los datos referentes a sus clientes, situación marital, estudios realizados por ellos, edad, niños en casa, adolescentes en casa, entre otros datos personales, anexado a también a la composición de su consumo en seis áreas específicas, consumo de pescado, carne, productos dulces, vino, frutas y productos de oro.

Para dicha empresa la clientela es frecuentemente la misma, con lo cual no le genera ingresos extras generar ofertas o promocionar productos en promoción, sino mas bien le interesa particularmente que los bienes consumidos por su clientes siempre estén disponibles en góndola, con lo cual necesariamente esta precisamente interesada en saber el consumo de su clientela con el fin de optimizar la satisfacción de sus clientes al momento de buscar lo que desea.

Supermercado Sanz nos contrata explicando su problemática, nos traspasa sus datos en formato .cvs y de inmediato nos ponemos a trabajar, armando un Análisis Exploratorio de Datos, limpiando y manipulando los datos para realizar el trabajo requerido.

El resultado del levantamiento del estudio de marketing de la empresa sobre una cantidad importante de personas, con datos de edad, nivel educacional, situación marital entre otras variables. Se busca determinar en función de los datos obtenidos de consumo, la caracterización de las personas más proclives a comprar los productos y cuanto compra dé cada tipo.

En el Dataset se observan un total de 29 columnas (considerando el ID) y un total de 2215 filas. Cada columa proporciona detalles específicos sobre los clientes y su consumo, pero a modo de resumen marcare y explicare las que considere más relevantes.

Year_Birth = Es un dato tipo entero, contiene los años de nacimiento de las personas encuestadas.

Education = Es un dato tipo texto, contiene el nivel maximo educativo alcanzado por las personas encuestadas.

Marital_Status = Es un dato tipo texto, contiene el estado marital de las personas encuestadas.

Income = Es un dato tipo entero, contiene el ingreso de las personas encuestadas.

Kidhome = Es un dato tipo entero, contiene el número de niños que viven en el mismo hogar que las personas encuestadas.

Teenhome = Es un dato tipo entero, contiene el número de adolescentes que viven en el mismo hogar que las personas encuestadas.

MntWines = Es un dato tipo entero, contiene el monto gastado en vinos en los ultimos 2 años.

MntFruits = Es un dato tipo entero, contiene el monto gastado en frutas en los ultimos 2 años.

MntMeatProducts = Es un dato tipo entero, contiene el monto gastado en carnes en los ultimos 2 años.

MntFishProducts = Es un dato tipo entero, contiene el monto gastado en pescado en los ultimos 2 años.

MntSweetProducts = Es un dato tipo entero, contiene el monto en dulces en los ultimos 2 años.

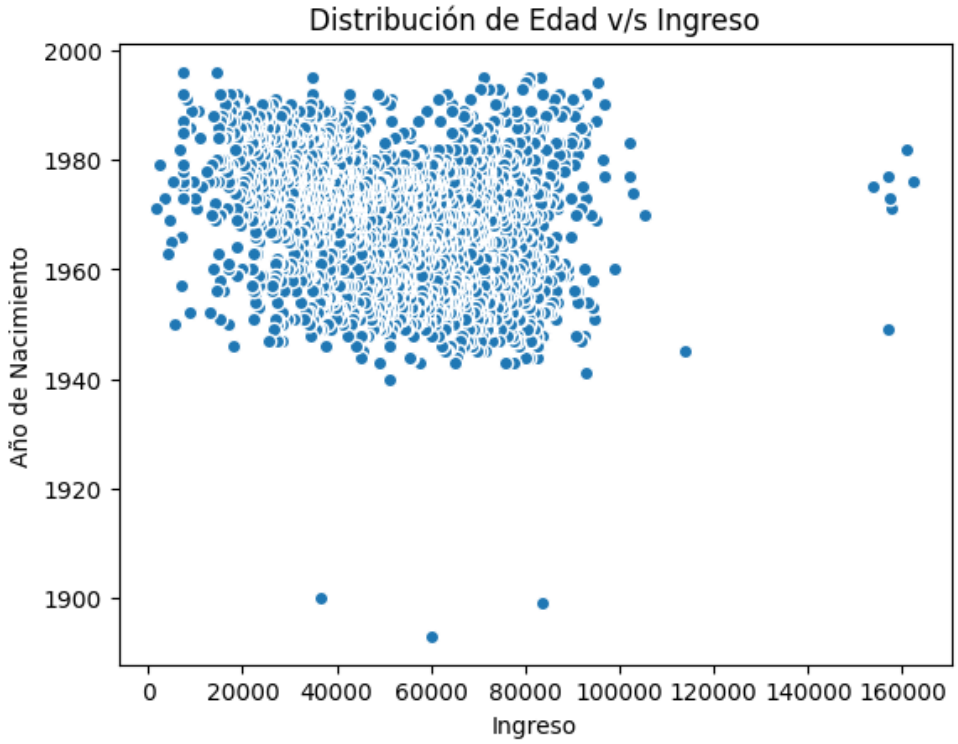
MntGoldProds = Es un dato tipo entero, contiene el monto en productos etiquetados como premium en los ultimos 2 años.

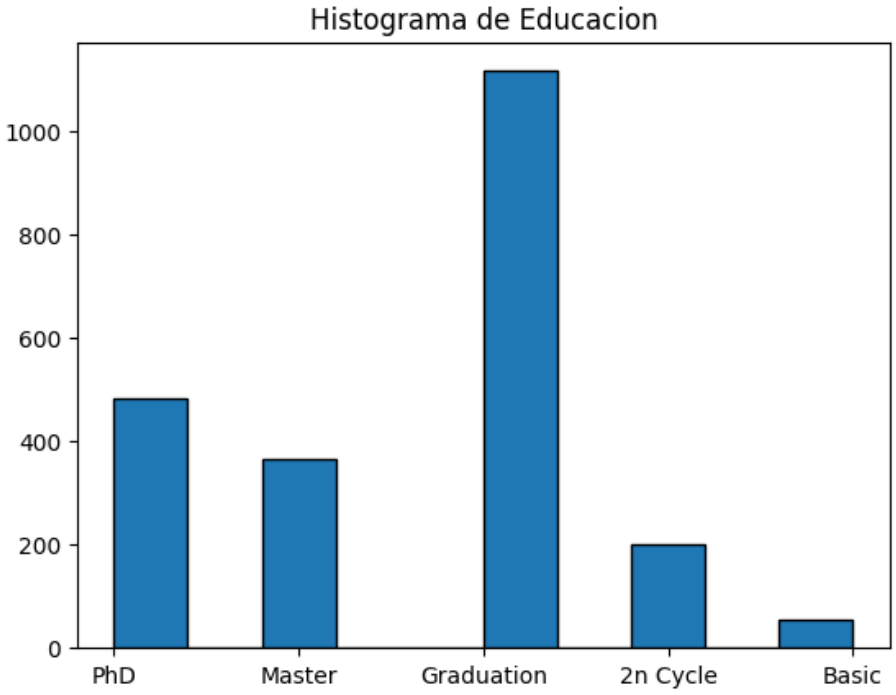
BREVE CONCLUSION SOBRE LOS CONSUMIDORES

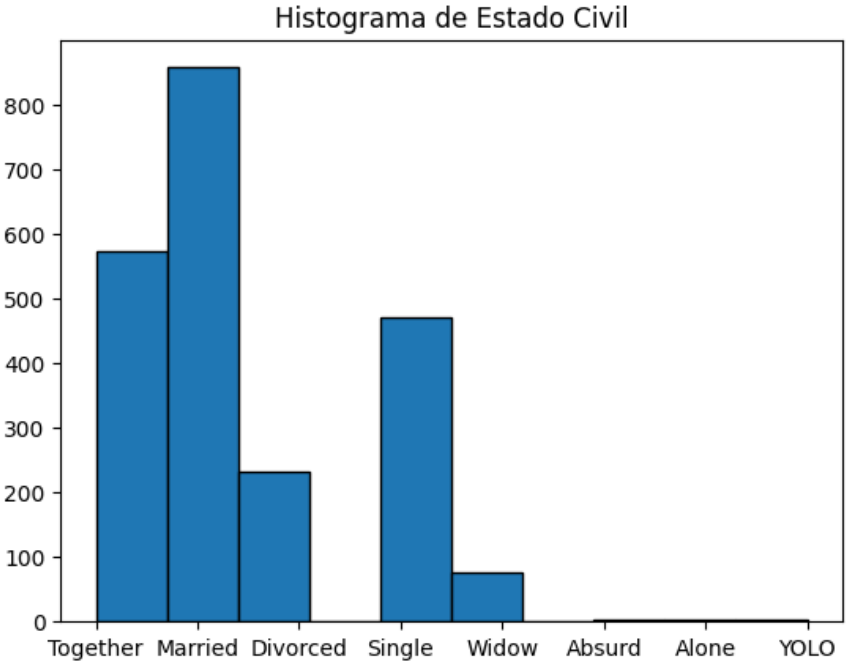
Dejo este link donde se puede ver este análisis en un Dashboard Dinamico

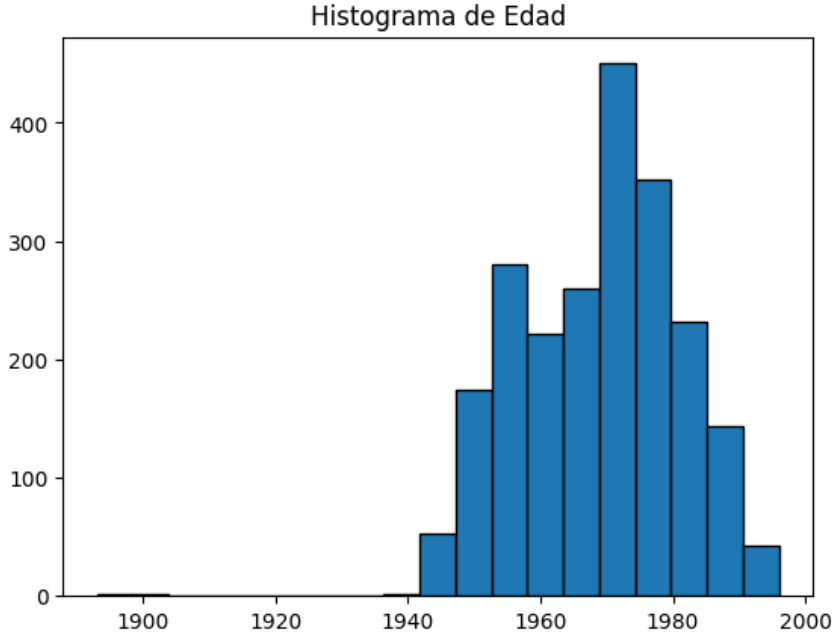
Tambien se puede visualizar el archivo Supermercado.html en este link

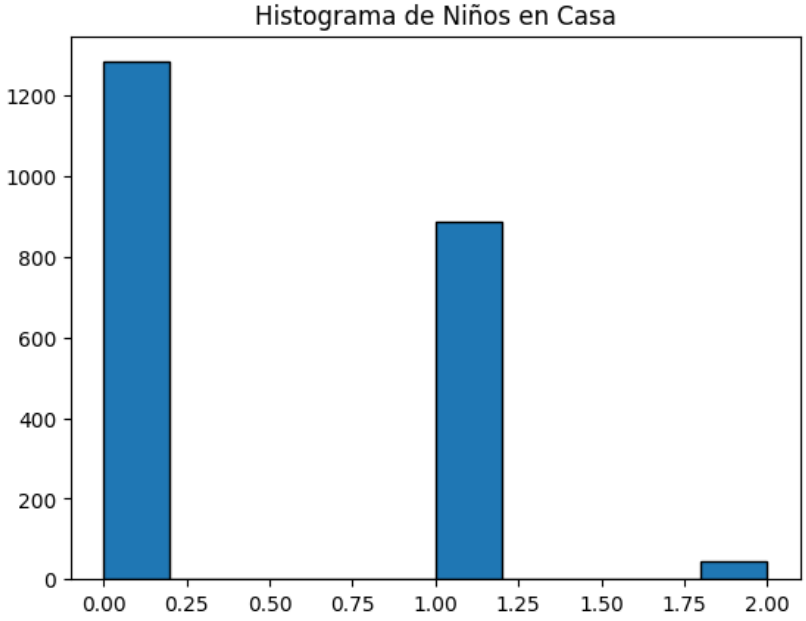
Los histograma nos demuestra cómo está distribuida nuestra población de acuerdo a su edad, estado civil, estudios y niños en casa. La mayor cantidad de nuestros consumidores nacieron entre 1950 y 1990, una buena parte de ellos están en pareja (Casados o conviviendo) y otra parte viviendos solos (solteros, viudos, divorciados). Por otro lado, la mayoría de los clientes o bien no tiene niños en casa o solo tiene un niño en casa. Por último los ingresos de los consumidores varían, pero en general están englobados entre 35000 y 70000 pesos.



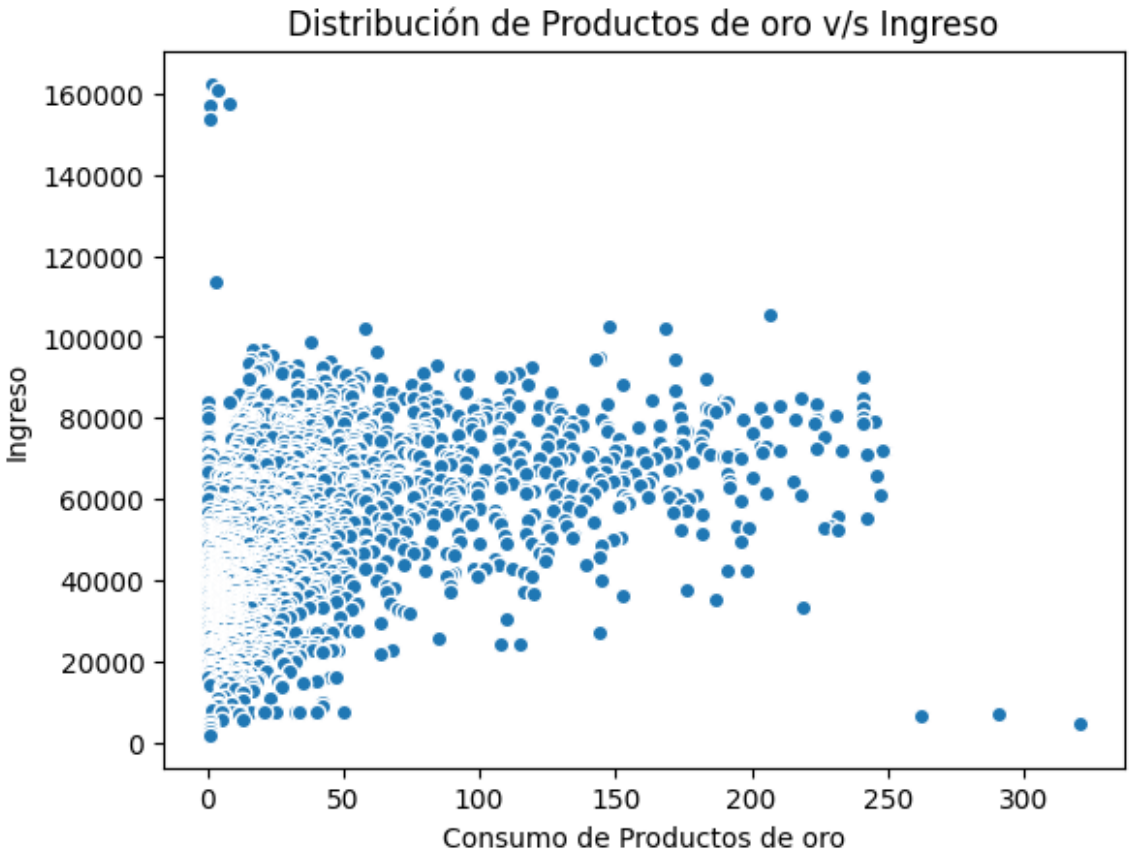


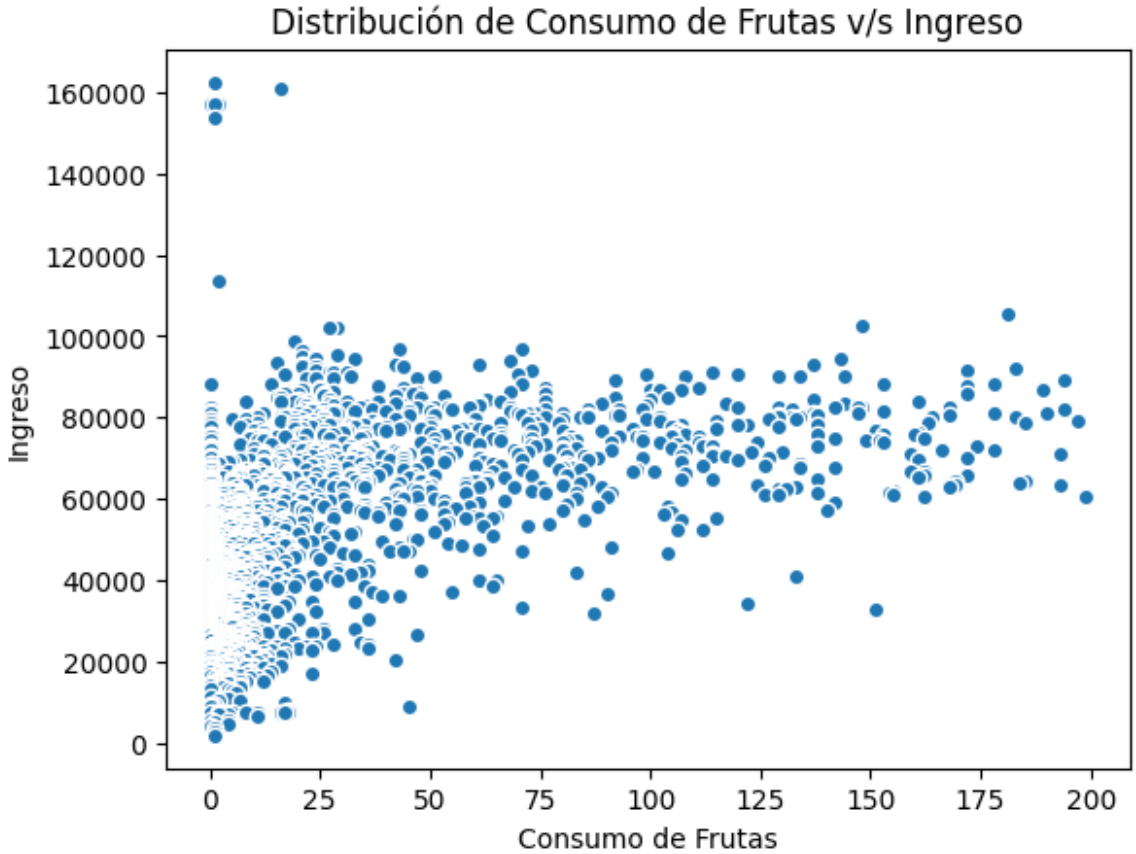

Histograma de Estado Civil
800
700
600
500
400
300
200
100
0
Together
Married
Divorced
Single
Widow
Absurd
Alone
YOLO

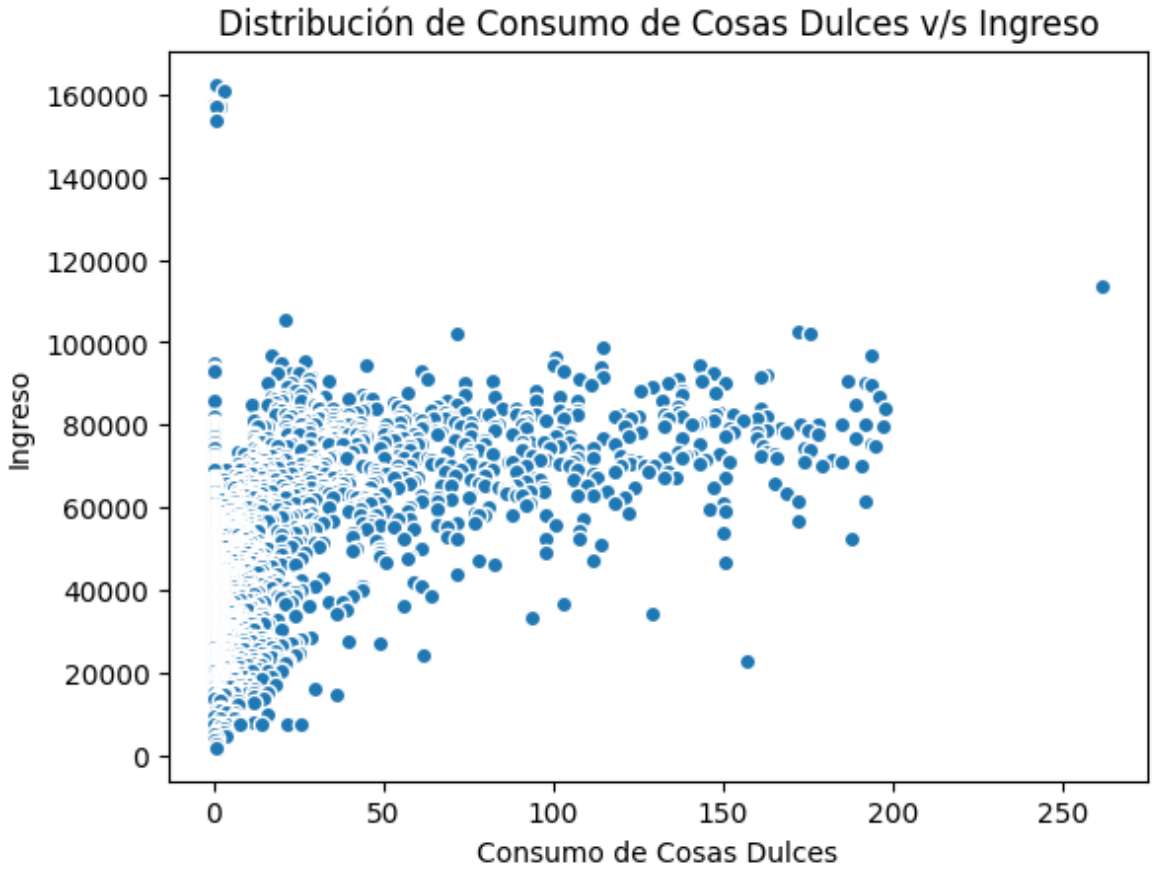

Histograma de Edad
400
300
200
100
0
1900
1920
1940
1960
1980
2000

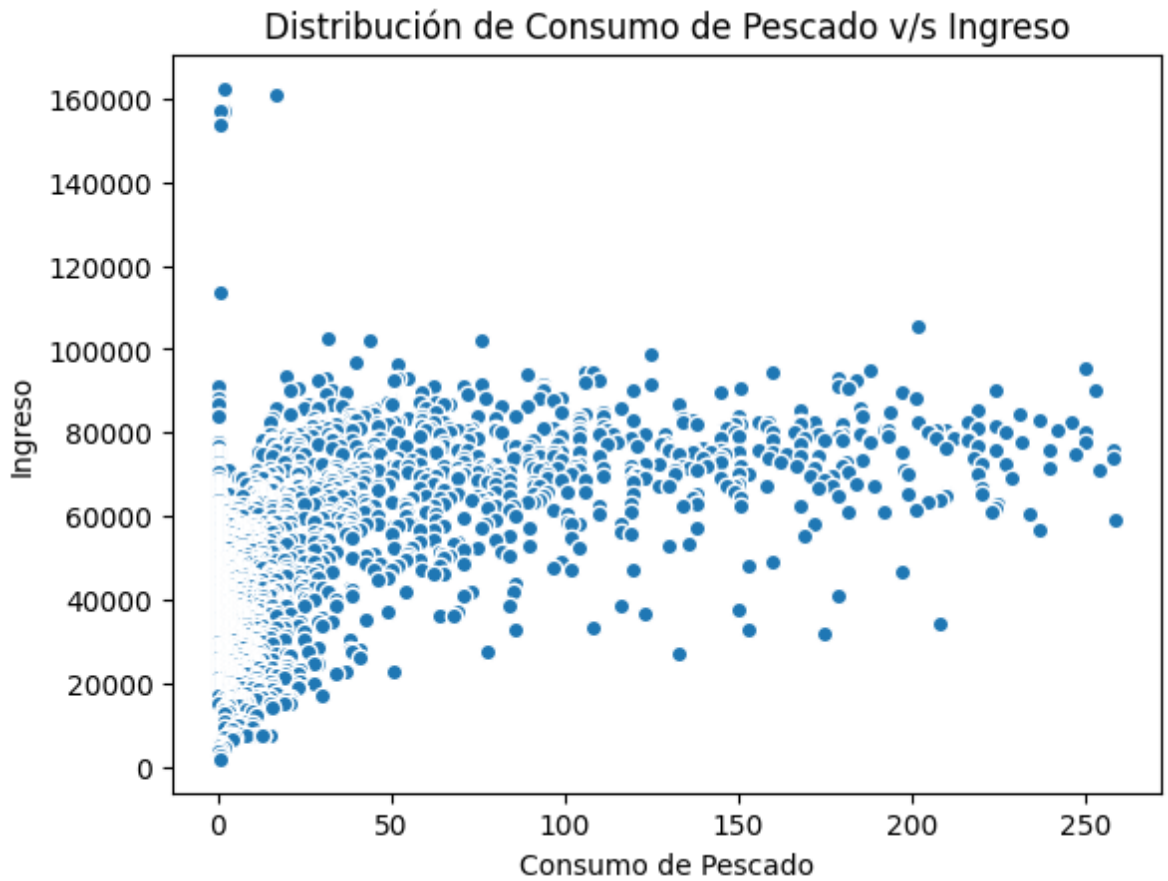

Histograma de Niños en Casa
1200
1000
800
600
400
200
0
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00

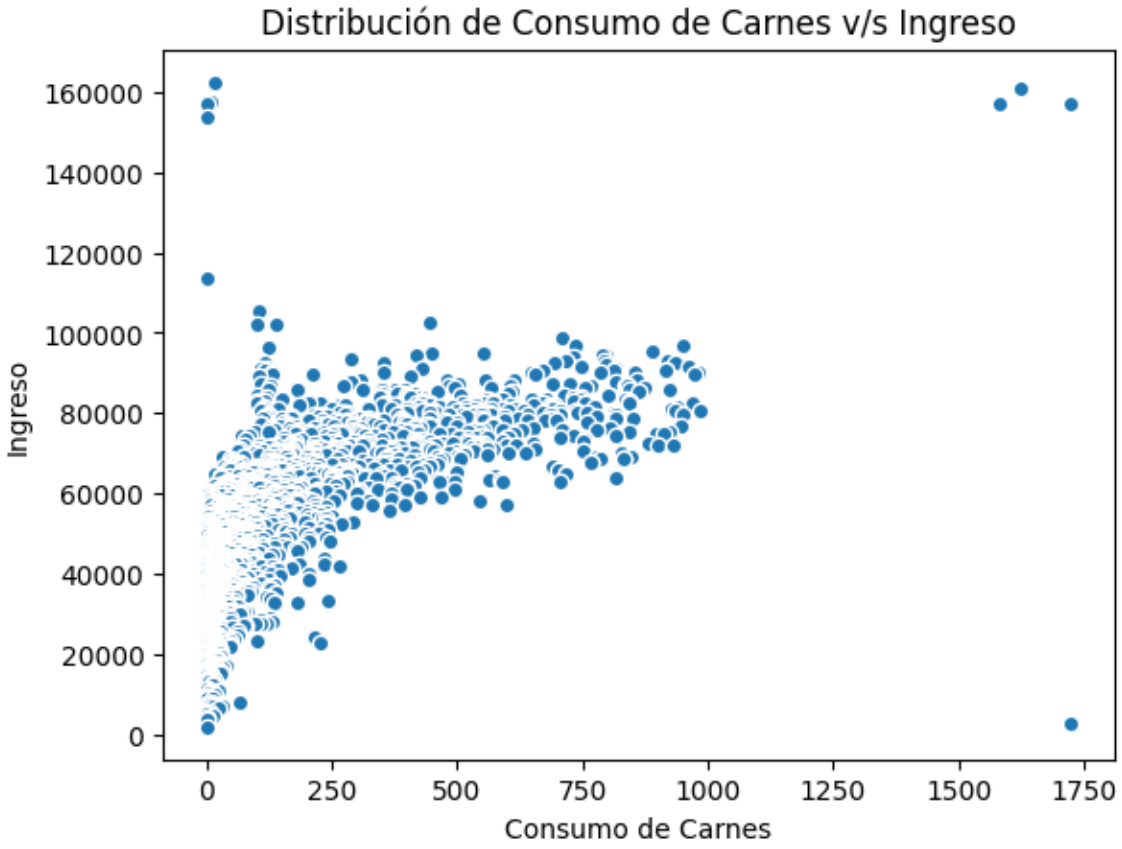
Con las visualizaciones siguientes se identifica que productos como el vino o la carne son muy consumidos y en particular la diferencia de cantidades entre los consumidores, el rango o la diferencia de lo mínimo y máximo de los que consumen en esos bienes es bastante grande. en cuando a los demás productos las cantidades consumidas es menor y más concentrado.

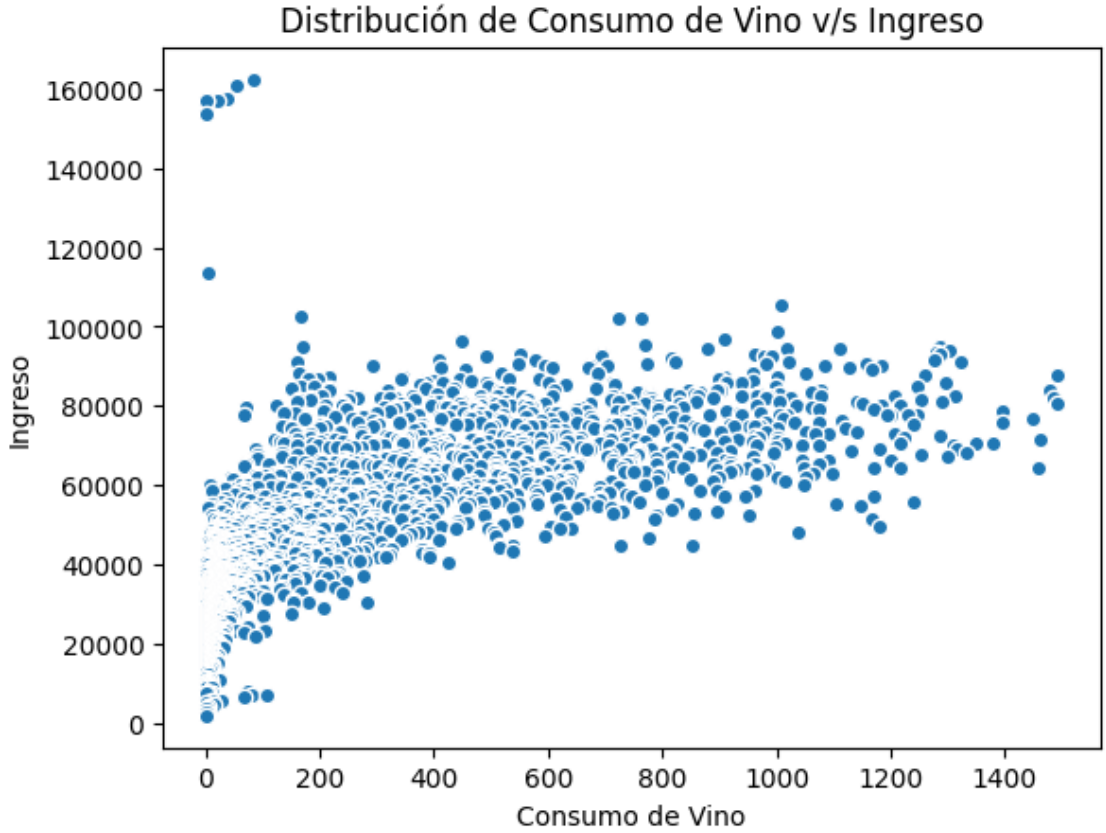





Distribución de Consumo de Cosas Dulces v/s Ingreso
Ingreso
160000
140000
120000
100000
80000
60000
40000
20000
0
0
50
100
150
200
250
Consumo de Cosas Dulces


Distribución de Consumo de Pescado v/s Ingreso
Ingreso
160000
140000
120000
100000
80000
60000
40000
20000
0
0
50
100
150
200
250
Consumo de Pescado


Distribución de Consumo de Carnes v/s Ingreso
Ingreso
Consumo de Carnes
0
20000
40000
60000
80000
100000
120000
140000
160000
0
250
500
750
1000
1250
1500
1750


Distribución de Consumo de Vino v/s Ingreso
Ingreso
Consumo de Vino
0
20000
40000
60000
80000
100000
120000
140000
160000
0
200
400
600
800
1000
1200
1400

A continuación nos ponemos a desarrollar un modelo de predicion de consumo, lo explicare en palabras para que luego la parte técnica se encargue de evaluar las técnicas realizadas.

La idea fue armar un modelo de machine learning, usando random forest para poder predecir el consumo en los distintos 6 segmentos de mercaderías, pescado, carne, vino, fruta, productos dulces y oro, pasadole como parámetros los datos del cliente, es decir según el ingreso, la cantidad de niños en casa, la cantidad de adolescentes en casa, la educación y el estado marital.

Para eso, se dividio el dataset original en tres partes, entrenamiento, validación y testeo, luego se sumó los datos de segmento de entrenamiento y validación.

También se tuvo que tratar dicha información de entrada, se hizo a traves de un Pipeline, donde se utilizo OneHotEncoder para cambiar variables categóricas en numéricas, también Binarizer y RobustScaler, para escalar o convertir valores numéricos en unos y ceros, entre otras herramientas.

Esto es la mejor precisión que conseguimos con este modelo

```
accuracy de Prediccion de consumo de Vino: 0.5077004779607011
accuracy de Prediccion de consumo de Frutas: 0.5432819968135953
accuracy de Prediccion de consumo de Carnes Rojas: 0.5087626128518322
accuracy de Prediccion de consumo de Pescado: 0.5443441317047265
accuracy de Prediccion de consumo de Productos Dulces: 0.5443441317047265
accuracy de Prediccion de consumo de Productos de Oro: 0.5114179500796601
```

Luego guardamos persistiendo el modelo en el archivo supermarket_pipeline.joblib para reutilizarlo posteriormente en caso de ser necesario…