El objetivo de este análisis era ir desgranando poco a poco los datos para ir sacando conclusiones.

Inicialmente quería dividir el proyecto en cinco partes y sacar conclusiones, pero la estrecha relación de unos con otros, donde uno se iba explicando mas a medida que se sacaban datos, hizo que separar los capítulos fuera más difícil, porque nunca quedaba cerrado del todo un capítulo, siempre se podía volver al anterior para justificar algún dato del nuevo.

Como análisis inicial hicimos un gráfico de barras para poder ver a números grandes el desempeño año tras año.

En vez de estar calculando columna por columna, para este cálculo usamos el método .pct_change, que calcula incrementos o decrementos porcentuales entre los datos de un dataframe.

Después hicimos otro grafico de barras agrupando por subcategoría para los datos de venta. Lo hicimos así y no por categoría porque daba mucha más información, de cara a un gráfico, cuantas más variables iba a dar datos mas diferenciados, de la otra manera, solo tendría 3 variables.

Para comprobar si la hipótesis en la que los pedidos medios son diferentes (H0) hicimos un ttest, quitando los outliers, tanto para unidades como para dólares.

En el caso de el pedido en valor decidimos quitar los pedidos que estaban más dispersos en el grafico de caja y bigotes, en cada año el umbral era diferente, pero tomé esa decisión porque tenía muchísimas líneas por encima del Q3, me iba a falsear el test si me quedaba solo con lo que estaban entre Q1 y Q3.

Para el pedido en unidades no quite ningún outlier porque los pedidos que estaban muy desviados suponían pocos pedidos sobre el total.

En el caso del beneficio, opte también por tratarlo a nivel subcategoría por la misma razón que en la venta, de hecho no podría haber localizado las pérdidas en el margen de la categoría mesas si no fuera por esto, ya que habría quedado diluida en su categoría (mobiliario)

Para el correlograma quise hacer variables que identificaran la venta, beneficio y descuento de cada año, es decir, vta_11, vta_12, ..., bco_11, bco_12, ..., d_11, d_12, ... pero no pude ponerlo todo junto para el correlograma, todo el rato me salía correlación perfecta, por lo que opté por hacer la correlación del total del periodo 2011-2014.

Aquí lo difícil, a parte de poner las variables juntas, fue interpretar qué significaban las diferentes columnas y como operaban entre ellas para poder tener datos que permitieran calcular correctamente otros datos, como por ejemplo, el descuento

en %, que si no lo calculábamos, daba 30% de media pero la realidad es que era un 44%

Así, decidí que la columna ‘sales’ era la venta neta, que ‘profit’ era el beneficio después de descuentos y con eso calculé ‘gross_sales’ y ‘discount_$’ y pude sacar el descuento total cada año.

En cuanto al número de referencias con descuento a pérdidas, simplemente fuimos haciendo una intersección de los valores únicos de cada año que tenían beneficio negativo.

Para finalizar, en el caso de los clientes y los pedidos, simplemente hicimos un conteo de los clientes y pedidos en cada año y lo pasamos a un gráfico de barras.

Lo más retador del código fue la cantidad de variables que se repetían, ya que teníamos que sacar para cada año la misma. Intenté hacerlas con alguna función donde tu le dijeras que año querías sacar pero me encontré con que no sabia como incluir ese input en el nombre de la variable para que me las generara iterando, por lo que el código es una consecución de repeticiones que parecen todas iguales.