Catalog Data – Mini Project 1

The file Catalog Dataset.zip contains a dataset in plain text format and two additional files – one that describes the data layout for each record in the data file, and one that describes the variables (data dictionary). Note that some fields are not included in the dataset (as noted in the data dictionary). The last two fields mentioned in the layout file (CR, LF) are the "carriage return" and "line feed" characters used to mark the end of the record. SAS will automatically handle this correctly. You do not need to read these characters separately.

The dataset is from a catalog company that mailed catalogs seasonally to existing customers, customers of subsidiary/affiliated companies, and customers reached via web advertising. The catalog was promoted through both direct mail and email. The data Set contains twelve years of data, through April 30, 2009. Information includes catalog orders, order source, quantity of items purchased, returns, payment information, and the zip Code of the purchaser. There is one record per order, with multiple orders per household. Orders with the same household are indicated with matching Household-ID numbers (one number per unique household.) The file contains 14,448 order records from 10,000 unique households. The accompanying data dictionary explains the data in more detail.

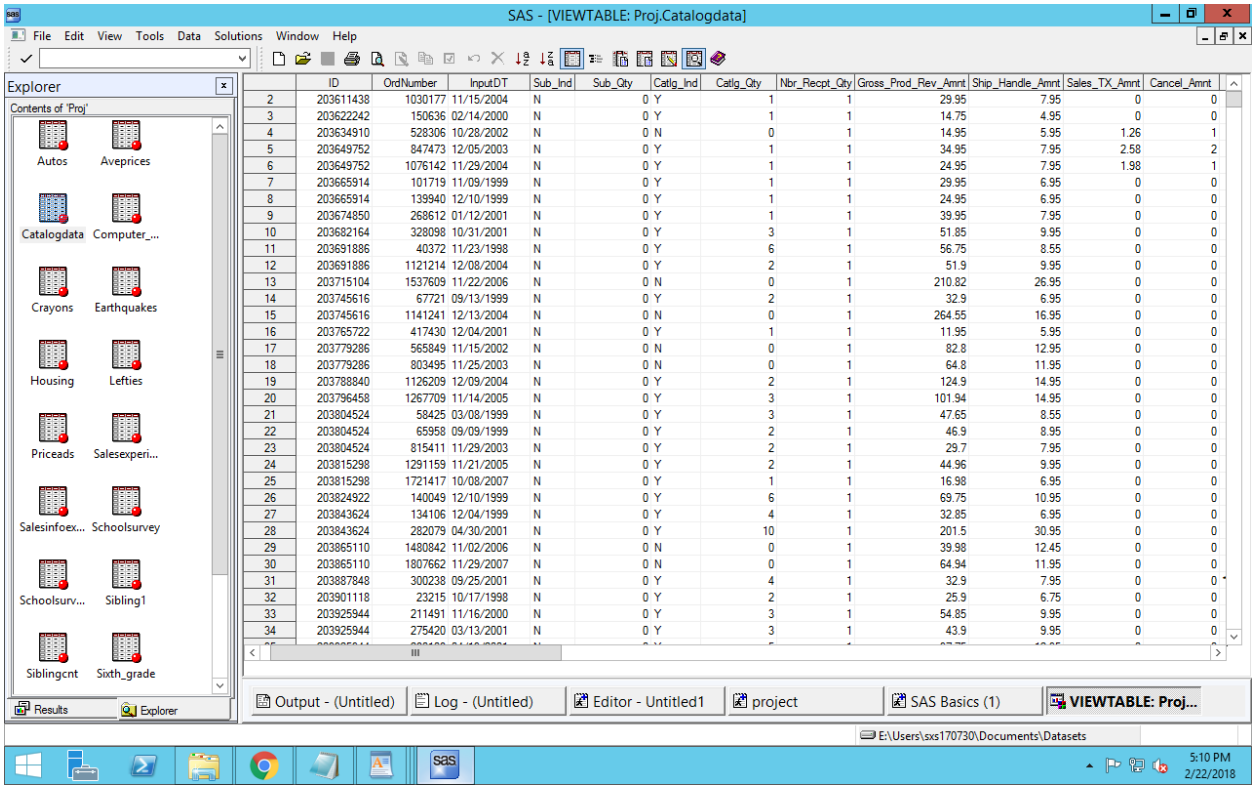
In addition, there is US zipcode data that provides the city, county and state details for each zipcode in the US. This can be useful to aggregate the data by geographic territories. There is also demographics data by zipcode from the US 2000 census data in the zip file US 2000 Census Data.zip. This can be used to understand how demographics affects catalog orders.

Initial Preparation

- Import the dataset into SAS (see the data organization file).
- Make sure that numbers are stored as numbers and dates are processed correctly. Perform additional processing if necessary to ensure that the fields are in the correct format.

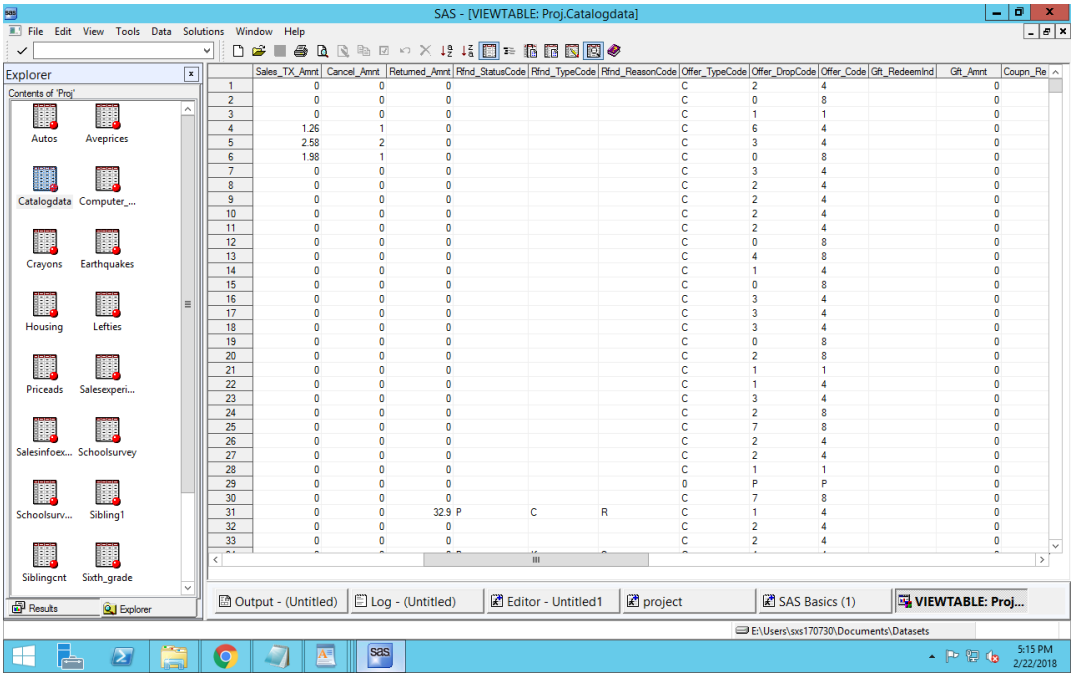
```
LIBNAME PROJ './';
DATA PROJ.CatalogData;
infile 'E:\Users\sxs170730\Documents\Datasets\Catalog-Data.txt';
input ID 1-10 OrdNumber 11-24 InputDT $ 25-34 Sub_Ind $ 35-35 Sub_Qty 36-42 Catlg_Ind $ 43-43
Catlg_Qty 44-50 Nbr_Recpt_Qty 51-64 Gross_Prod_Rev_Amnt 65-72 Ship_Handle_Amnt 73-80
Sales_TX_Amnt 81-88
Cancel_Amnt 81-86 Returned_Amnt 97-104 Rfnd_StatusCode $ 131-131 Rfnd_TypeCode $ 132-132
Rfnd_ReasonCode $ 133-133
Offer_TypeCode $ 134-134 Offer_DropCode $ 135-135 Offer_Code $ 136-136 Gft_RedeemInd $ 139-
139 Gft_Amnt 140-147
Coupn_RedeemInd $ 148-148 Coupn_Amnt 149-156 Gft_Ind $ 157-157 Pmnt_TypCode 158-161
Pmnt_CategoryCode $ 162-162
Pmnt_StatusCode $ 163-163 Ord_TypCode $ 164-164 Ancllry_Ind $ 165-165 Ancllry_Qty 166-179
Addl_ChrgAmnt 180-187
```

Addnl_ChrgCode $ 188-189 Web_Ind $ 190-190 Web_Qty 191-197 Write_OffAmnt 198-205 Div_Code $ 206-207 Ind_ID 208-234
Rank_InputDT 235-241 Cust_Zipcode 242-246;
RUN;

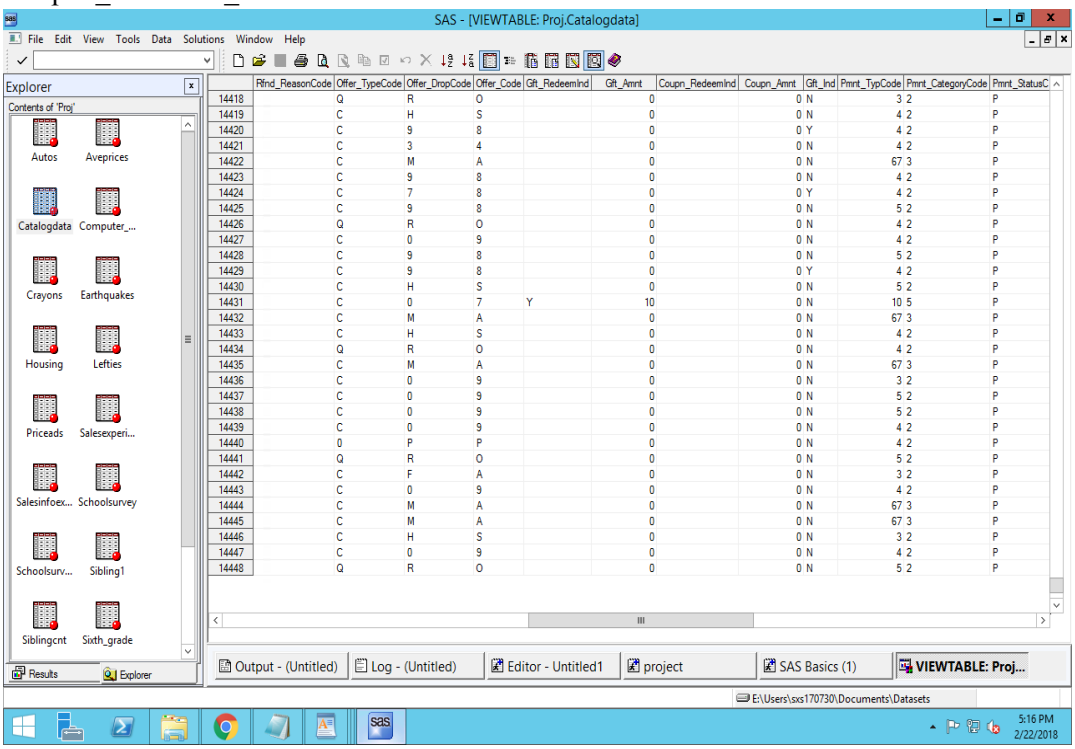


	ID	OrdNumber	InputDT	Sub_Ind	Sub_Qty	Catlg_Ind	Catlg_Qty	Nbr_Recpt_Qty	Gross_Prod_Rev_Amnt	Ship_Handle_Amnt	Sales_TX_Amnt	Cancel_Amnt
2	203611438	1030177	11/15/2004	N	0	Y	1	1	29.95	7.95	0	0
3	203622242	150636	02/14/2000	N	0	Y	1	1	14.75	4.95	0	0
4	203634910	528306	10/28/2002	N	0	N	0	1	14.95	5.95	1.26	1
5	203649752	847473	12/05/2003	N	0	Y	1	1	34.95	7.95	2.58	2
6	203649752	1076142	11/29/2004	N	0	Y	1	1	24.95	7.95	1.98	1
7	203665914	101719	11/09/1999	N	0	Y	1	1	29.95	6.95	0	0
8	203665914	139940	12/10/1999	N	0	Y	1	1	24.95	6.95	0	0
9	203674850	268612	01/12/2001	N	0	Y	1	1	39.95	7.95	0	0
10	203682164	328098	10/31/2001	N	0	Y	3	1	51.85	9.95	0	0
11	203691886	40372	11/23/1998	N	0	Y	6	1	56.75	8.55	0	0
12	203691886	1121214	12/08/2004	N	0	Y	2	1	51.9	9.95	0	0
13	203715104	1537609	11/22/2006	N	0	N	0	1	210.82	26.95	0	0
14	203745616	67721	09/13/1999	N	0	Y	2	1	32.9	6.95	0	0
15	203745616	1141241	12/13/2004	N	0	N	0	1	264.55	16.95	0	0
16	203765722	417430	12/04/2001	N	0	Y	1	1	11.95	5.95	0	0
17	203779286	565849	11/15/2002	N	0	N	0	1	82.8	12.95	0	0
18	203779286	803495	11/25/2003	N	0	N	0	1	64.8	11.95	0	0
19	203788840	1126209	12/09/2004	N	0	Y	2	1	124.9	14.95	0	0
20	203796458	1267709	11/14/2005	N	0	Y	3	1	101.94	14.95	0	0
21	203804524	58425	03/08/1999	N	0	Y	3	1	47.65	8.55	0	0
22	203804524	65958	09/09/1999	N	0	Y	2	1	46.9	8.95	0	0
23	203804524	815411	11/29/2003	N	0	Y	2	1	29.7	7.95	0	0
24	203815298	1291159	11/21/2005	N	0	Y	2	1	44.96	9.95	0	0
25	203815298	1721417	10/08/2007	N	0	Y	1	1	16.98	6.95	0	0
26	203824922	140049	12/10/1999	N	0	Y	6	1	69.75	10.95	0	0
27	203843624	134106	12/04/1999	N	0	Y	4	1	32.85	6.95	0	0
28	203843624	282079	04/30/2001	N	0	Y	10	1	201.5	30.95	0	0
29	203865110	1480842	11/02/2006	N	0	N	0	1	39.98	12.45	0	0
30	203865110	1807662	11/29/2007	N	0	N	0	1	64.94	11.95	0	0
31	203887848	300238	09/25/2001	N	0	Y	4	1	32.9	7.95	0	0
32	203901118	23215	10/17/1998	N	0	Y	2	1	25.9	6.75	0	0
33	203925944	211491	11/16/2000	N	0	Y	3	1	54.85	9.95	0	0
34	203925944	275420	03/13/2001	N	0	Y	3	1	43.9	9.95	0	0

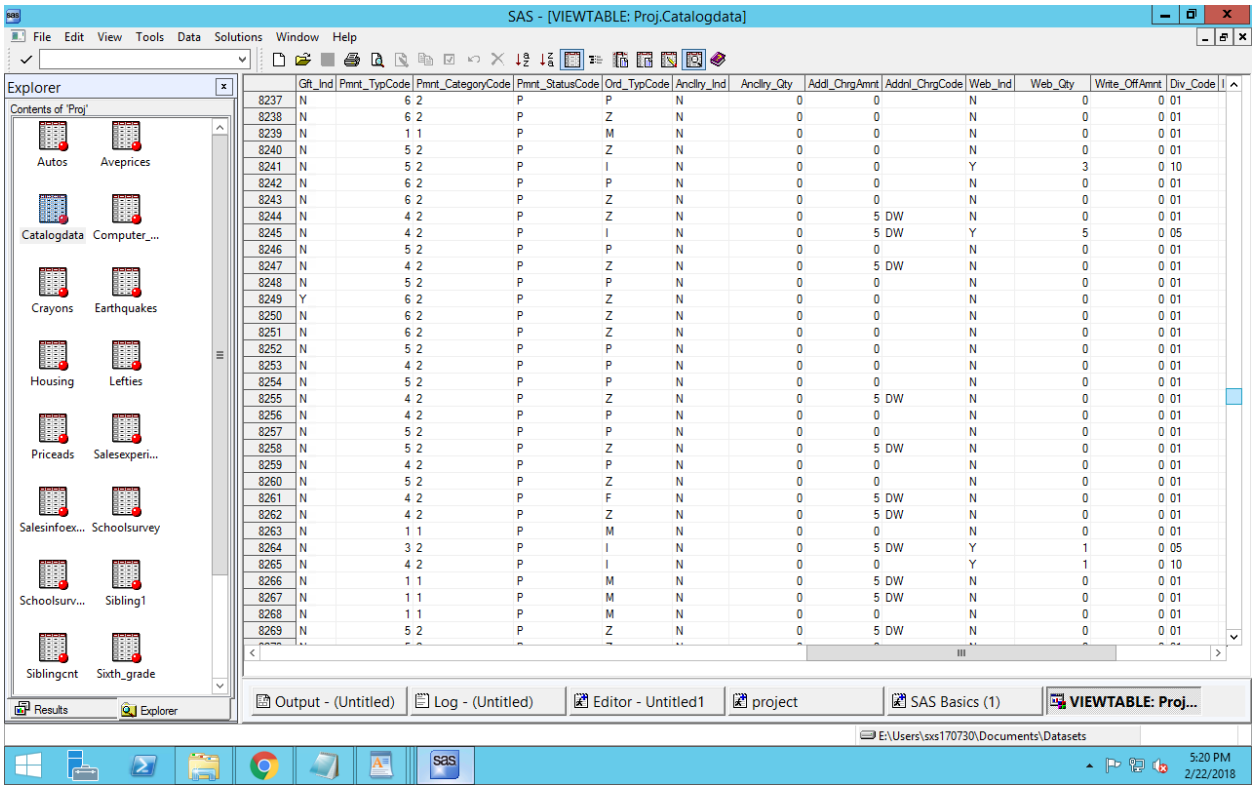
- Examine the data in each field of the dataset. Look for missing data. The objective is to understand the organization and meaning of the data better.
  - Coupon redeemed indicator→ all blanks
  - Refund status code, Refund Type code, Refund Reason code → there are blanks if no status is mentioned



- Gift_Certificate_Redeemed_indicator → only one field is populated as 'Y'. Rest of the fields are blank. We can replace the blank fields with 'N' if there is any requirement.
- Coupon_redeemed_indicator → same as above but all fields are blank in this case.



- Additional charges amount → a lot of blanks are present



	Gft_Ind	Pmnt_TypCode	Pmnt_CategoryCode	Pmnt_StatusCode	Ord_TypCode	Ancllry_Ind	Ancllry_Qty	Addl_ChrgAmnt	Addnl_ChrgCode	Web_Ind	Web_Qty	Write_OffAmnt	Div_Code
8237	N	6	2	P	P	N	0	0		N	0	0	01
8238	N	6	2	P	Z	N	0	0		N	0	0	01
8239	N	1	1	P	M	N	0	0		N	0	0	01
8240	N	5	2	P	Z	N	0	0		N	0	0	01
8241	N	5	2	P	I	N	0	0		Y	3	0	10
8242	N	6	2	P	P	N	0	0		N	0	0	01
8243	N	6	2	P	Z	N	0	0		N	0	0	01
8244	N	4	2	P	Z	N	0	5	DW	N	0	0	01
8245	N	4	2	P	I	N	0	5	DW	Y	5	0	05
8246	N	5	2	P	P	N	0	0		N	0	0	01
8247	N	4	2	P	Z	N	0	5	DW	N	0	0	01
8248	N	5	2	P	P	N	0	0		N	0	0	01
8249	Y	6	2	P	Z	N	0	0		N	0	0	01
8250	N	6	2	P	Z	N	0	0		N	0	0	01
8251	N	6	2	P	Z	N	0	0		N	0	0	01
8252	N	5	2	P	P	N	0	0		N	0	0	01
8253	N	4	2	P	P	N	0	0		N	0	0	01
8254	N	5	2	P	P	N	0	0		N	0	0	01
8255	N	4	2	P	Z	N	0	5	DW	N	0	0	01
8256	N	4	2	P	P	N	0	0		N	0	0	01
8257	N	5	2	P	P	N	0	0		N	0	0	01
8258	N	5	2	P	Z	N	0	5	DW	N	0	0	01
8259	N	4	2	P	P	N	0	0		N	0	0	01
8260	N	5	2	P	Z	N	0	0		N	0	0	01
8261	N	4	2	P	F	N	0	5	DW	N	0	0	01
8262	N	4	2	P	Z	N	0	5	DW	N	0	0	01
8263	N	1	1	P	M	N	0	0		N	0	0	01
8264	N	3	2	P	I	N	0	5	DW	Y	1	0	05
8265	N	4	2	P	I	N	0	0		Y	1	0	10
8266	N	1	1	P	M	N	0	5	DW	N	0	0	01
8267	N	1	1	P	M	N	0	5	DW	N	0	0	01
8268	N	1	1	P	M	N	0	0		N	0	0	01
8269	N	5	2	P	Z	N	0	5	DW	N	0	0	01

Data Inspection and Cleaning

- The data dictionary describes how orders, returns, cancellations, refunds, shipping etc. are handled. Is the data consistent with this description? If not, what are the discrepancies? How will you handle the discrepancies (if any)?
  - Household ID (ID) can have multiple orders (represented by unique order_number). Households can be divided as per zipcodes and do an analysis taking the demographic data
  - The magazine subscription quantity is used with subscription indicator (='Y')
  - **Gross Product Revenue Amount:**
    Pre 1/25/2007 → the total amount purchased on this order. This total does not include shipping/handling and taxes. Coupon amounts and additional charges (to the company) are NOT removed from GPR - these charges must be removed manually if needed. The return and cancel amounts are subtracted from the GPR.
    Post 1/25/2007 → This field will be changed to include all shipping, handling and tax amounts. Nothing will be subtracted from the GPR. Any returns, cancels, coupons, etc. will have to be subtracted manually.
  - **Shipping and handling amount:**
    Post 1/25/2007→ CORDR.GROSS_PRODUCT_REVENUE_AMOUNT will be changed to include all shipping, handling and tax amounts. Nothing will be subtracted from the GPR. Any returns, cancels, coupons, etc. will have to be subtracted manually.
    Pre-1/25/2007:→ This amount is NOT included in GROSS_PRODUCT_REVENUE_AMOUNT.

- **Sales tax amount:**
  Post 1/25/2007→ CORDR.GROSS_PRODUCT_REVENUE_AMOUNT will be changed to include all shipping, handling and tax amounts. Nothing will be subtracted from the GPR. Any returns, cancels, coupons, etc. will have to be subtracted manually.
  Pre-1/25/2007: This amount is NOT included in CORDR.GROSS_PRODUCT_REVENUE_AMOUNT.
- **Cancel Amount:**
  Details: For each CANCELLED line item on an order (line item action code = C), the number of units cancelled times the retail price is determined. This field is the sum of all cancelled line items on the order.
  As of 1/25/2007: CORDR.GROSS_PRODUCT_REVENUE_AMOUNT will be changed to include all shipping, handling and tax amounts. Nothing will be subtracted from the GPR. Any returns, cancels, coupons, etc. will have to be subtracted manually.
  Pre-2007: The cancel amount for this order. This amount is subtracted from GROSS_PRODUCT_REVENUE_AMOUNT.
- **Returned Amount:**
  The returned amount for this order.
  Details: For each RETURNED line item on an order (line item action code = R), the number of units returned times the retail price is determined. This field is the sum of all returned line items on the order. (Note this does not include EXCHANGED line items.)
  Post 1/25/2007: GROSS_PRODUCT_REVENUE_AMOUNT will be changed to include all shipping, handling and tax amounts. Nothing will be subtracted from the GPR. Any returns, cancels, coupons, etc. will have to be subtracted manually.
  Pre-2007: The returned amount for this order. This amount is subtracted from CORDR.GROSS_PRODUCT_REVENUE_AMOUNT.
- **Refund Amount:**
  The refund amount for this order.
  Details: This is the dollar amount issued back to the customer due returns, exchanges, sell outs, and cancellations of line items. For example, a customer pays with a check and then we sell out of their item. That would spawn a refund of the check amount to the customer.
  Post 1/25/2007: GROSS_PRODUCT_REVENUE_AMOUNT will be changed to include all shipping, handling and tax amounts. Nothing will be subtracted from the GPR. Any returns, cancels, coupons, etc. will have to be subtracted manually.
  Pre-2007: The refund amount for this order. This amount is NOT subtracted from CORDR.GROSS_PRODUCT_REVENUE_AMOUNT. The return and cancel amounts, however, ARE subtracted from CORDR.GROSS_PRODUCT_REVENUE.
- The division code is also important

- The data dictionary mentions that refunds, cancellations, returns and other order information are handled differently depending on the transaction date. Update the dataset such that the data has the same consistent interpretation regardless of the date.
- Are there other variables in the dataset that require inspection and cleaning (to ensure consistent interpretation or avoid missing data)? Are there other variables that are interdependent and need to be checked for consistency? Are there data fields that you can / should drop?

Exploratory Analysis

- What are the different ways in which you can categorize orders? What variables are useful and reliable for this purpose?
- What are different ways in which you can categorize customers?
- Are there any interesting patterns or trends that you see in the data? Over time? Across geographies?

Developing Marketing Insights

- Brainstorm and generate a list of marketing ideas / questions / insights relevant to this dataset. These marketing ideas / questions/ insights should help a marketing manager take some marketing actions to improve the business. So for item on your list, you need to be clear how or why it would be useful for a marketing manager.
  REMEMBER: Not every analysis or hypothesis test you can perform with the data is useful to know for a manager. So think carefully about why a certain question should be interesting or important.
- Frame each item on the list as a question that you can answer using the data.
- For three of these, conduct the necessary analysis to answer the question using the data. Consider whether there is more than one way to pose the question and/ or conduct the analysis.

Presenting Insights

- Prepare a 10-min presentation to discuss the three insights that you obtained from the data.
  REMEMBER: Not every statistically significant result, is actually important for a manager. Think carefully about whether or how that piece of insight will affect marketing decisions. For e.g., even if household income has a significant in fact on sales, if the actual magnitude of the effect is not that large (given the range of possible household income), then this finding is not of much consequence.
- In your presentation, describe each item, explain why it is important or interesting for marketing, then present your analysis and conclusions, and finish by making recommendations for specific marketing actions that the firm can take.

Deliverables

- Project Report
    - What changes / assumptions you made while cleaning the data

    - List of marketing ideas you identified. Why these are interesting or important for amarketing manager. Exploratory analysis, if any, that led to these ideas.
    - The analysis, results and conclusions for three of the ideas that you chose for your presentation
    - Include all SAS code in the Appendix.
- Project Presentation as discussed above.