

Pharmacy Dataset (AB Dataset)

Project Description:

The dataset contains transactional data of pharmacy store, which is then done further divided into product master, pharmacy master, POS transaction. The data set contain one excel workbook with 7 sheets of data. The data in each sheet is mention below:

1. PHRMCY MASTER: Pharmacy Master with set of Pharmacy IDs (surrogate keys), de-identified Pharmacy names, State Cd & Zip 3 Cd (1098 records)
2. PROD MASTER: Product Master contains product description, major category id, category code, sub category code and segment code (189,053 records)
3. MAJOR PROD CAT: Major Category Codes and its description (15 records)
4. PROD CAT: Product Category Codes and category description (63 records)
5. PROD SUB CAT: Product Sub-Category Codes and its description (246 records)
6. PROD SEG: Product Segment Codes and segment description (1005 records)
7. POS TRANS: Point-of-Sales transactions with Sales Dates of for six months, from 2016-01-01 through 2016-06-30 (915,744 records)

Pre-Processing of Data

FOR EDA

We had 7 sheets in excel files like Product Master, POS master, Category master etc so we had import individual sheet into Enterprise guide.

The sheets are then joined using product number, pharmacy number, major category, category, sub-category and segment numbers.

Once all the individual sheets were merged, we had combine all the files into a single SAS format Dataset which can be used for Enterprise miner and further analysis.

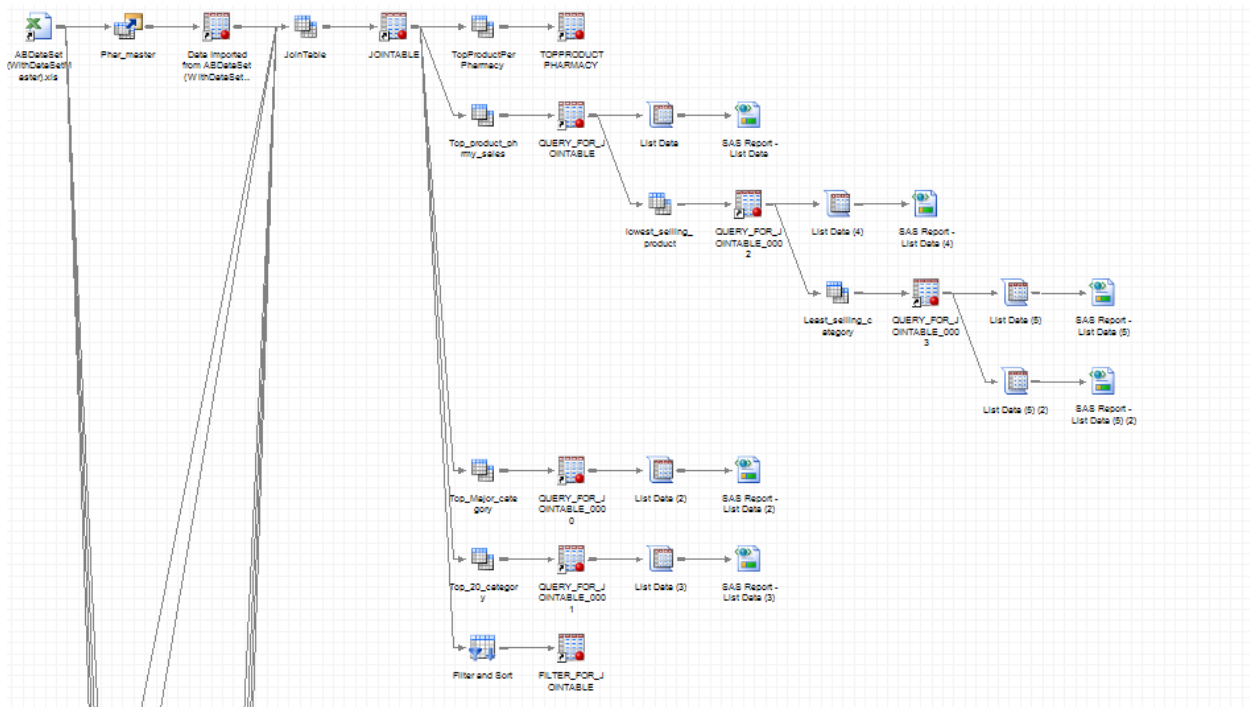


Figure 1: Guide Process Flow

For Forecasting:

We have done forecast on sales quantity and Zip code. Although initially we want to perform forecasting on Product number and Sales quantity, but SAS Forecast Studio has limitation of 1000 level only., we can't perform Forecasting on Product number.

The Hierarchical Data model is used with TOP to BOTTOM approach.

The Hierarchy used here is Major Product Category/ Category / Sub Category / Segment

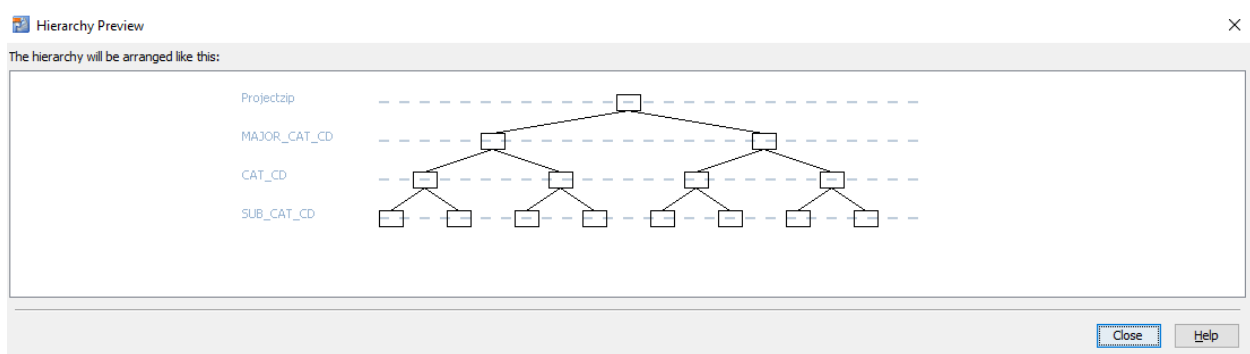



Figure 2: Hierarchy Structure

The Date format is not in SAS Date but in numeric form.

SEG_CD	SEG_DESC	EXT_SLS_...	SLS_QTY	SLS_DTE_...	date
530053105311	CANDY	1.6	2	20160608	20160608
580058505857	OTHER CARDS	3.49	1	20160117	20160117
600060306032	FACIAL TISSUE	3.49	1	20160117	20160117
130013401345	THICKENERS...	24.79	1	20160119	20160119
370037803787	THERAPEUTI...	6.29	1	20160127	20160127

Figure 3: SAS Date Format

We have converted numeric format to SAS Date format using INPUT and PUT Function.

1 of 2 Build an advanced expression 

Enter an expression:

```
INPUT(PUT(1.SLS_DTE_NBR, 8.), YYMMDD10.)
```

Figure 4: Date Format Formula

For ZIP code the format is in text field we have convert to numeric field for analysis in Forecast.

Enter an expression:

```
INPUT(1.ZIP_3_CD, 3.)
```

Figure 5: ZIP code Formula





 SLS_DTE_...	 date	 ZIP_3_CD	 NumericZip
20160608	20160608	071	71
20160117	20160117	109	109
20160117	20160117	025	25
20160119	20160119	115	115
20160127	20160127	180	180
20160127	20160127	180	180
20160106	20160106	046	46
20160201	20160201	077	77

Figure 6: ZIP Code Numeric Format

Sales Quantity as Dependent variable and Zip code as independent variables.

For Sales Quantity Sum of values is chosen and ZIP code Total number of values is chosen

Variable Roles In Your Data

Variable	Role	Aggregation	Accumulation	Usage in System-Generat...
SLS_QTY	Dependent	Sum of values	Sum of values	
NumericZip	Independent	Total number of values	Total number of values	Try to use
EXT_SLS_AMT	None			
SLS_DTE_NBR	None			

Figure 7: Variables Roles

Explanatory Data Analysis

Top 25 selling products across all pharmacy:

Top_25_selling_products_among_all_phrmcy

Row number	Total_Sales	PROD_NBR	PROD_DESC	PHRMCY_NBR
1	21046	92000000000	GENERICQS1ITEM	61520549788616420
2	13526	90800000000	VITAMINS/SUPPLEMENTS	61520549788616420
3	12057	1820025008	01820025008	4416100399456673861
4	11098	40003000753	MIDWEST FASTENER	9201518331233084207
5	10230	90400000000	DME SALES	61520549788616420
6	8022	99100000066590000000006	CANDY OPEN DEPARTMENT	6991356705459241502
7	7629	99100000120141000000010	CANDY & BEVERAGE	4416100399456673861
8	7431	90400000000	DME SALES	9201518331233084207
9	6228	90600000000	CARDS	7999580510622165484
10	6080	99100000066590000000006	CANDY OPEN DEPARTMENT	2090966447983750424
11	5831	90400000000	DME SALES	5464072734544936345
12	5472	99100000770330000000003	MONEY ORDER	4416100399456673861
13	5291	91100000000	CIGARETTES	5464072734544936345
14	4453	991000000819890000000040	STAMPS	4416100399456673861
15	3454	98650000000000944904150	WHEELED WALKER BRA/SEAT	2487426938853675539
16	2878	90900000000	STAMPS	9201518331233084207
17	2836	91800000000	STOMACH	9201518331233084207
18	2719	2820000357	MARLBORO BOX	6991356705459241502
19	2703	991000001797420000000050	SYRINGE & NEEDLE	4416100399456673861
20	2496	90300000000	DME	5464072734544936345
21	2464	90300000000	DME	3009693108150153253
22	2380	91000000000	LOTTO	7003025686214903268
23	2358	98650000000000185441572	STRUTZ PRO	2487426938853675539
24	2341	98650000000000273630700	FRAME	3216540913770909647
25	2287	8085322022	DUBRA VODKA 200	4416100399456673861

Figure 8: Top 25 Selling Products

Top Selling Major Category

Top_selling_Major_Category

Row number	MAJOR_CAT_CD	MAJOR_CAT_DESC	total_sales
1	5228	HEALTH CARE	419863
2	4371	EDIBLES	182153
3	6137	GENERAL MERCHANDISE	150992
4	9687	GREETING CARDS	140057
5	3391	HOME HEALTH CARE	85497
6	7392	PERSONAL CARE	82243
7	7020	BEAUTY	47447
8	2343	DIABETES	11123
9	1941	PHOTO	9953
10	5068	MISC	2639

Figure 9: Top Selling Major Category

Top 20 selling Categories

Top_20_Selling_Category			
Row number	CAT_CD	CAT_DESC	Total_Sales
1	5800	GREETING CARDS & OTHER ASSOCIATED MANUFACTURER ITEMS	140057
2	0700	COLD & ALLERGY	125793
3	5300	CONFECTIONS	94614
4	5600	FOOD & BEVERAGES	87539
5	4100	VITAMINS/DIETARY SUPPLEMENTS	69397
6	0100	PAIN RELIEF	62811
7	0300	DIGESTIVE HEALTH	56530
8	9300	HOME HEALTH CARE	56268
9	6300	MISC GENERAL MERCHANDISE	51291
10	2500	FIRST AID	50475
11	7100	TOBACCO	39254
12	3500	ORAL CARE	34565
13	3700	SKIN CARE	27528
14	1900	EYE & EAR CARE	18148
15	6000	HOUSEHOLD PRODUCTS	15025
16	2700	FOOT CARE	13292
17	0500	BABY CARE	13213
18	2900	HAIR CARE	12210
19	2300	FEMININE CARE	11781
20	9400	INCONTINENCE	11724

Figure 10: Top 20 Selling Products

10 MOST RETURN PRODUCTS

Most_Return_Products				
Row number	Total_Sales	PROD_NBR	PROD_DESC	PHRMCY_NBR
1	-8	90300000000	DME	9201518331233084207
2	-3	73588290395	DG CARD BIRTHDAY	4416100399456673861
3	-3	7467645311	MUELLER KNEE BRACE ADJUSTABLE ONE SIZE	4325879497219228989
4	-2	31011900238	BAUSCH+LOMB SENS EYES+SALINE 12OZ	7055791495512548111
5	-2	3320018370	ARM&H DENTL CARE PASTE ADV CLN BSP 6.3OZ	1841593177282919617
6	-2	31254717140	BENADRYL ITCH RELIEF STICK 0.47OZ	536719232234314822
7	-2	30904110231	POVIDINE IODINE 10% OINTMENT 1OZ MAJOR	536719232234314822
8	-2	30193708050	CONTOUR TEST STRIP 50CT	3241946627864485090
9	-2	7565600066	NECKLACE NITE GLOW 20 36/CT JD OBS	4416100399456673861
10	-2	90300000000	DME	1571515534661781510

Figure 11: Most Return Products

TOP NON-SELLING PRODUCTS, CATEGORIES

NON_SELLING_PRODUCTS						
Row number	Total_Sales	MAJOR_CAT_CD	MAJOR_CAT_DESC	CAT_CD	CAT_DESC	PROD_DESC
1	0	2343	DIABETES	1100	DIABETES CARE	BD ULTRAFINE 12.7MM 30GX0.5CC 100CT
2	0	2343	DIABETES	1100	DIABETES CARE	ONE TOUCH ULTRA 2 METER
3	0	2343	DIABETES	1100	DIABETES CARE	PRODIGY NO CODE STRIP MEDI 50CT
4	0	2343	DIABETES	1100	DIABETES CARE	GNP TRUERSULT GLUC MONTOR KIT
5	0	2343	DIABETES	1100	DIABETES CARE	FREESTYLE LITE TEST STRIP 50CT
6	0	2343	DIABETES	1100	DIABETES CARE	MEDICOOL DIA-PAK DAYMATE
7	0	2343	DIABETES	1100	DIABETES CARE	INSULIN COOLING CASE
8	0	2343	DIABETES	1100	DIABETES CARE	02571566711
9	0	2343	DIABETES	1100	DIABETES CARE	ONE TOUCH ULTRA MINI METER SILVER
10	0	2343	DIABETES	1100	DIABETES CARE	JOB SOCK CREW W/MD 8-15M
11	0	2343	DIABETES	1100	DIABETES CARE	DIASTIX REAGENT STRIPS 50CT
12	0	2343	DIABETES	1100	DIABETES CARE	DIADERM FOOT CREAM REJUVENATING
13	0	3391	HOME HEALTH CARE	1400	PHYSICAL FITNESS & EXERCISE EQUIPMENT	PEDLAR EXER CHRM
14	0	3391	HOME HEALTH CARE	1400	PHYSICAL FITNESS & EXERCISE EQUIPMENT	PULLEY EXER SET
15	0	3391	HOME HEALTH CARE	1500	HEALTH SUPPORTS	THERMOPHORE HEAT MAX MD 14X14

Figure 12: Non-selling Products

Forecast Results

The Summary of Forecast Results for Major category, category and Sub Category with zip codes

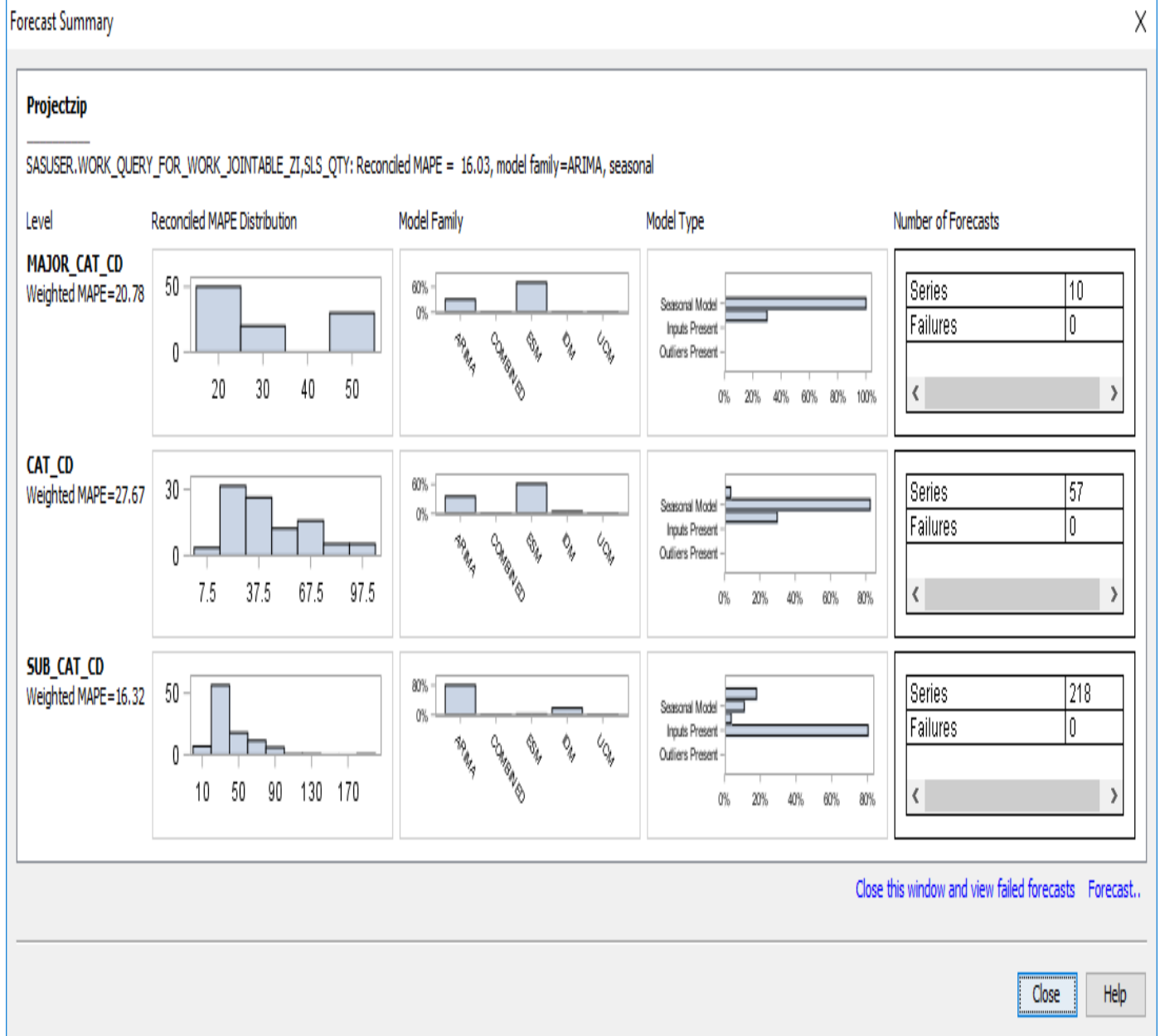


Figure 13: Overall Summary

For overall hierarchy the best model is 7020 major category(Cosmetic) with 5400(Beauty) with MAPE value of 1.30

Filter: All				
(286 of 286 series)				
MAJOR_...	CAT_CD	SUB_CA...	MAPE /	Rec. MAPE
7020	7300	*	0	3.26
7020	7300	73007310	0	3.26
5228	7800	*	0	9.24
5228	7800	78007800	0	9.24
3391	9300	93009370	0	17.77
7020	5400	54005440	0	30.77
7020	3900	39003910	1.30	39.83
5228	0100	01000110	1.90	18.29
3391	9000	90009070	2.28	23.69
5228	0700	07000710	2.76	23.10

Figure 14: Best Hierarchy

The Forecasting of beauty products show a upward and downward trends.

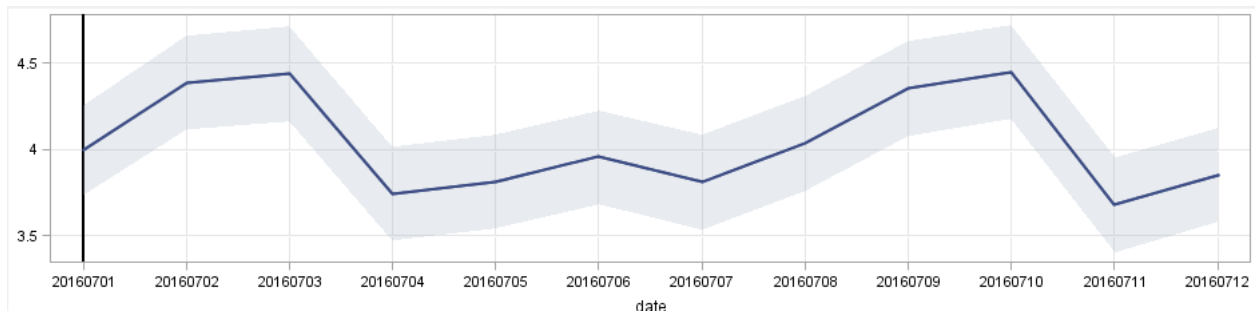


Figure 15: Forecast on Beauty Products

The best model is ARIMA model with has lowest MAPE values:

<input checked="" type="checkbox"/> Active series Forecast Model: HPF2_626(automatic selection) Set this model as forecast model			
Model	Type	Read-Only	MAPE
Generated ARIMA Model (HPF2_626)	Generated	Yes	1.30
Generated ARIMA Model (HPF2_627)	Generated	Yes	1.51
Generated Smoothing Model (HPF2_628)	Generated	Yes	57.42

Figure 16: Beauty Product Model

The Parameter estimates are shown below. Sales quantity are not significant with respect to previous sales.

Parameter Estimates					
Component	Parameter	Estimate	Standard Error	t Value	Approx Pr > t
SLS_QTY	CONSTANT	-0.03139	0.02552	-1.23	0.2245
SLS_QTY	MA1_7	-0.02861	0.14645	-0.20	0.8459
SLS_QTY	AR1_1	-0.29623	0.24618	-1.20	0.2345
NumericZip	SCALE	1.01734	0.0076039	133.79	<.0001

Figure 17: Parameters Estimates

Major category forecast

The major category which has best model is 5228 Health Care which has lowest MAPE value of 15.06

MAJOR_...	MAPE /	Rec. MAPE
5228	15.06	17.23
7392	15.32	16.91
7020	19.69	21.64
3391	21.33	23.89
6137	22.00	25.12
4371	22.07	24.66
9687	33.59	34.99
5068	45.59	48.30
2343	47.51	49.52
1941	49.38	50.94

Figure 18: Major Category Model

The forecast model doesn't predict the linear trends in sales of health care products. There are upward and downward trends in sales of previous data.

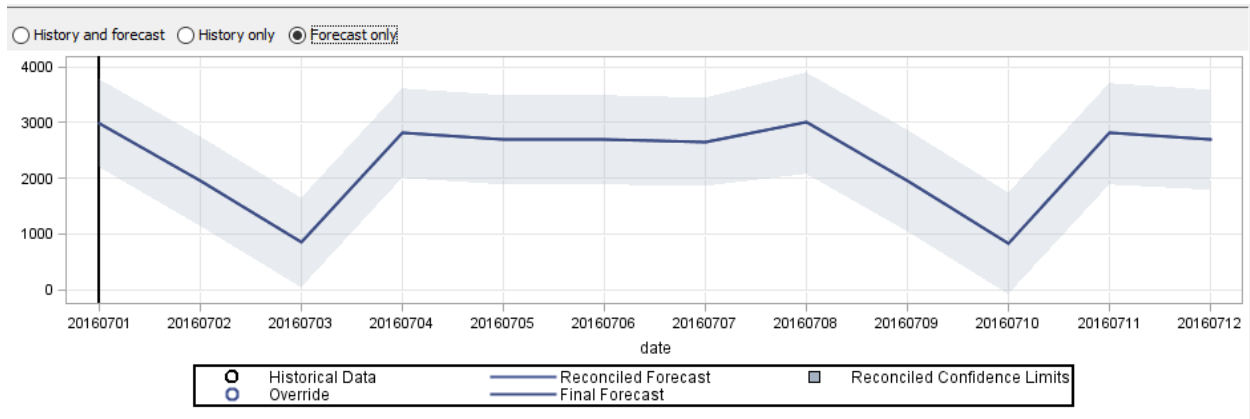


Figure 19: Forecast of Health Care Products

The model selected here is ARIMA model with MAPE value of 15.06

Active series Forecast Model: HPF0_15(automatic selection) Set this model as forecast model			
Model	Type	Read-Only	MAPE
Generated ARIMA Model (HPF0_15)	Generated	Yes	15.06
Generated Smoothing Model (HPF0_16)	Generated	Yes	17.91

Figure 20: Model selection of Health Care

The parameter estimates are given below. The Sales of previous 7 days are highly significant here.

Component	Parameter	Estimate	Standard Error	t Value	Approx Pr > t
SLS_QTY	MA1_7	0.45498	0.06772	6.72	<.0001

Figure 21: Parameter Estimates of Health Care

Regression

The Regression is used to see the correlation between the Categories of Products and Zip Code. We have used three regression model covering major category, category and zip code. The data set is filter to sales quantity having 1-10 values only as SAS was unable to take more then 512 levels of target variables.

The data is partition into 70:30 ratio with random sampling.

Meta data is used to modify the input to three different Regression models. The Overall process flow is given below.

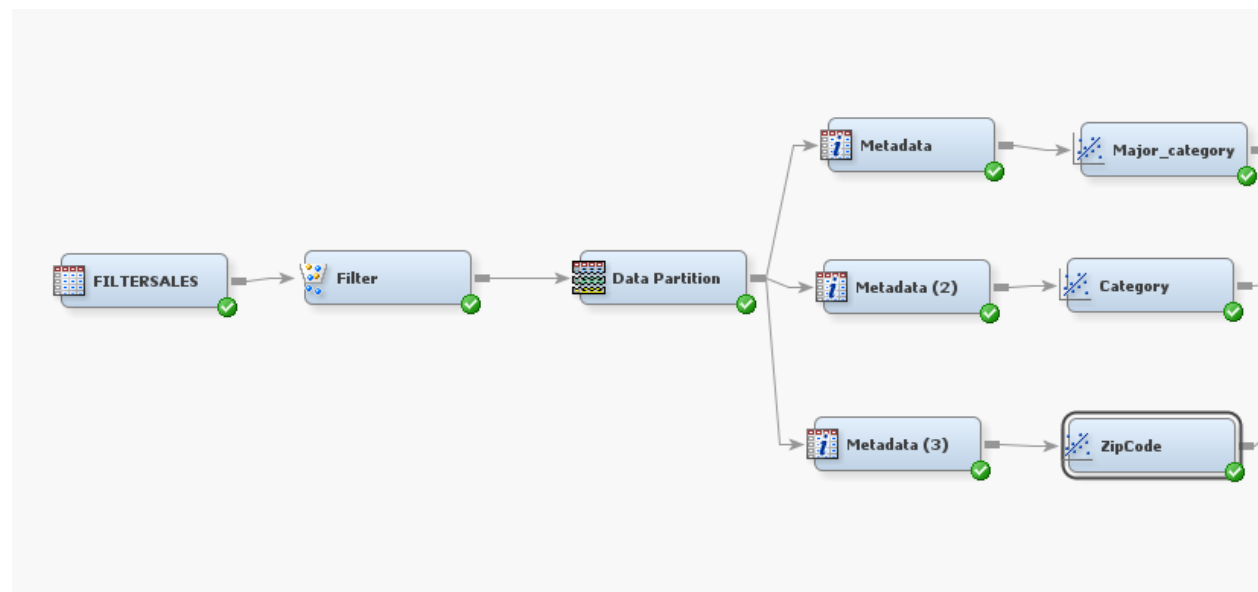


Figure 22: Regression process Model

Regression model with major category as input

The result of regression shows major category 3391(health care) and 4371(Edible) as significant for most of sales quantity. The likelihood is positive for 4371 edible products with sales quantity. Whereas for health care likelihood is negative suggestion negative correlation

185	MAJOR_CAT_CD	3391	10	1	0.2454	0.4735	0.27	0.6042	1.278
186	MAJOR_CAT_CD	3391	9	1	0.7188	0.5614	1.64	0.2004	2.052
187	MAJOR_CAT_CD	3391	8	1	0.3097	0.4797	0.42	0.5185	1.363
188	MAJOR_CAT_CD	3391	7	1	0.8099	0.4974	2.65	0.1035	2.248
189	MAJOR_CAT_CD	3391	6	1	-0.2691	0.4480	0.36	0.5480	0.764
190	MAJOR_CAT_CD	3391	5	1	0.2038	0.4476	0.21	0.6488	1.226
191	MAJOR_CAT_CD	3391	4	1	-0.1729	0.4439	0.15	0.6970	0.841
192	MAJOR_CAT_CD	3391	3	1	-0.2889	0.4428	0.43	0.5142	0.749
193	MAJOR_CAT_CD	3391	2	1	-0.5637	0.4421	1.63	0.2023	0.569
194	MAJOR_CAT_CD	3391	1	1	-0.8956	0.4419	4.11	0.0427	0.408
195	MAJOR_CAT_CD	4371	10	1	2.5889	0.6585	15.46	<.0001	13.315
196	MAJOR_CAT_CD	4371	9	1	2.1909	0.7241	9.15	0.0025	8.943
197	MAJOR_CAT_CD	4371	8	1	2.1054	0.6637	10.06	0.0015	8.210
198	MAJOR_CAT_CD	4371	7	1	2.0772	0.6790	9.36	0.0022	7.982
199	MAJOR_CAT_CD	4371	6	1	2.2086	0.6434	11.78	0.0006	9.103
200	MAJOR_CAT_CD	4371	5	1	1.8317	0.6445	8.08	0.0045	6.244
201	MAJOR_CAT_CD	4371	4	1	1.7667	0.6423	7.57	0.0059	5.852
202	MAJOR_CAT_CD	4371	3	1	1.3649	0.6419	4.52	0.0335	3.915
203	MAJOR_CAT_CD	4371	2	1	1.2028	0.6416	3.51	0.0608	3.329
204	MAJOR_CAT_CD	4371	1	1	0.5660	0.6415	0.78	0.3776	1.761
205	MAJOR_CAT_CD	5228	10	1	0.3576	0.4099	0.76	0.3829	1.430

Figure 23: MLE for Categories

Regression model with category as input

The result of category as independent variable and sales as dependent variables, doesn't show any significant relationship between sales and category of products

Analysis of Maximum Likelihood Estimates								
Parameter	SLS_QTY	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp (Est)	
Intercept	10	1	3.1429	3.8512	0.67	0.4145	23.171	
Intercept	9	1	2.0782	4.0064	0.27	0.6040	7.990	
Intercept	8	1	2.9202	3.8711	0.57	0.4506	18.544	
Intercept	7	1	2.6289	3.9044	0.45	0.5007	13.859	
Intercept	6	1	4.0452	3.8051	1.13	0.2877	57.124	
Intercept	5	1	4.1234	3.8017	1.18	0.2781	61.770	
Intercept	4	1	5.1318	3.7834	1.84	0.1750	169.326	
Intercept	3	1	6.0037	3.7769	2.53	0.1119	404.930	
Intercept	2	1	7.7504	3.7730	4.22	0.0400	999.000	
Intercept	1	1	10.1866	3.7723	7.29	0.0069	999.000	
CAT_CD	0100	10	-0.4516	4.1013	0.01	0.9123	0.637	
CAT_CD	0100	9	-0.1320	4.2637	0.00	0.9753	0.876	
CAT_CD	0100	8	-0.0786	4.1178	0.00	0.9848	0.924	
CAT_CD	0100	7	-0.1312	4.1545	0.00	0.9748	0.877	
CAT_CD	0100	6	-0.2945	4.0478	0.01	0.9420	0.745	
CAT_CD	0100	5	-0.2527	4.0440	0.00	0.9502	0.777	
CAT_CD	0100	4	-0.1427	4.0236	0.00	0.9717	0.867	
CAT_CD	0100	3	0.1108	4.0164	0.00	0.9780	1.117	
CAT_CD	0100	2	0.3587	4.0123	0.01	0.9288	1.432	
CAT_CD	0100	1	1.0369	4.0116	0.07	0.7960	2.820	
CAT_CD	0300	10	-0.0971	4.0112	0.00	0.9807	0.907	
CAT_CD	0300	9	-0.1184	4.1739	0.00	0.9774	0.888	
CAT_CD	0300	8	-0.0245	4.0315	0.00	0.9952	0.976	
CAT_CD	0300	7	0.0696	4.0653	0.00	0.9863	1.072	
CAT_CD	0300	6	0.1568	3.8628	0.00	0.9688	1.167	

Figure 24: MLE For Category & Sales

Regression model with ZIP code as input

The Model for ZIP code and Sales quantity doesn't generate any significant result. Hence, we say that there doesn't exist a relationship between Zip code and sales quantity. The Intercept does represent some significant values

Analysis of Maximum Likelihood Estimates

Parameter	SLS_QTY	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)
Intercept	10	1	3.2025	1.0689	8.98	0.0027	24.593
Intercept	9	1	1.9968	1.1166	3.20	0.0737	7.365
Intercept	8	1	2.8272	1.0783	6.87	0.0087	16.897
Intercept	7	1	2.5033	1.0896	5.28	0.0216	12.222
Intercept	6	1	4.0415	1.0568	14.62	0.0001	56.913
Intercept	5	1	4.0701	1.0569	14.83	0.0001	58.563
Intercept	4	1	5.2698	1.0504	25.17	<.0001	194.372
Intercept	3	1	6.0628	1.0489	33.41	<.0001	429.573
Intercept	2	1	7.8113	1.0479	55.57	<.0001	999.000
Intercept	1	1	10.2738	1.0477	96.16	<.0001	999.000
ZIP_3_CD 010	10	1	0.2309	1.3294	0.03	0.8621	1.260
ZIP_3_CD 010	9	1	0.3078	1.3829	0.05	0.8238	1.360
ZIP_3_CD 010	8	1	1.4374	1.3328	1.16	0.2808	4.210
ZIP_3_CD 010	7	1	-0.0290	1.3577	0.00	0.9829	0.971
ZIP_3_CD 010	6	1	0.9654	1.3138	0.54	0.4625	2.626
ZIP_3_CD 010	5	1	-0.2703	1.3174	0.04	0.8375	0.763
ZIP_3_CD 010	4	1	-0.1478	1.3085	0.01	0.9101	0.863
ZIP_3_CD 010	3	1	-0.6590	1.3069	0.25	0.6141	0.517
ZIP_3_CD 010	2	1	-0.9238	1.3053	0.50	0.4791	0.397
ZIP_3_CD 010	1	1	-0.8780	1.3049	0.45	0.5010	0.416
ZIP_3_CD 016	10	1	-0.0904	3.2360	0.00	0.9777	0.914
ZIP_3_CD 016	9	1	-0.0248	3.3795	0.00	0.9941	0.975
ZIP_3_CD 016	8	1	-0.0618	3.2644	0.00	0.9849	0.940
ZIP_3_CD 016	7	1	-0.0207	3.2961	0.00	0.9950	0.980
ZIP_3_CD 016	6	1	-0.1876	3.1996	0.00	0.9532	0.829
ZIP_3_CD 016	5	1	-0.1962	3.1990	0.00	0.9511	0.822
ZIP_3_CD 016	4	1	-0.0632	3.1755	0.00	0.9841	0.939
ZIP_3_CD 016	3	1	-0.2214	3.1713	0.00	0.9443	0.801

Recommendation

1. The top selling categories such as Cold and Allergy, Vitamins, Pain Relief (Health Care) greeting cards should always be in stock
2. Top selling products such as Vitamins/Supplements, Midwest Fastener, Generic items, Cigarettes, Candy should always be in stock
3. The categories which are least sold or not sold at all such as Diabetes Care, Physical fitness and exercise equipment should not be carried as they result in loss of revenue.
4. Most of the products which are returned such as Birthday card, Mueller Knee brace, Bausch & Lomb sense eyes +saline, analysis should be made as to why these products are returned frequently and whether they should be kept in stock
5. Lowest forecasted Photo category should only be kept on shelves for limited period
6. Health care product category have good forecast with forecasting of 7 days relevant to next sales. This could help to predict the future procurement of health care products