Group 1

Data Mining and Business Intelligence

Ekaterina Burkhanova, Neehar Namjoshi, Pooja Shah

Table of Contents

Introduction of Dataset

Date	Store	Dept		Weekly_Sa	Type	Size	Temperat	Fuel_Price	CPI	Unemploy	IsHoliday	Year	Month	Week	max	min	mean	median	std	Total_MarkDown
5/2/19	1	1	5/2/19	24924.5	A	151315	42.31	2.572	211.0964	8.106	0	2019	2	5	57592.12	14537.37	22513.32	18535.48	9854.349	0
5/2/19	9	97	5/2/19	668.48	B	125833	38.01	2.572	214.6555	6.415	0	2019	2	5	766.93	-9.92	372.6556	371.05	290.9547	0
5/2/19	9	85	5/2/19	693.87	B	125833	38.01	2.572	214.6555	6.415	0	2019	2	5	2512.14	110.56	876.6294	824.04	307.4361	0
5/2/19	8	80	5/2/19	8654.6	A	155078	34.14	2.572	214.4715	6.299	0	2019	2	5	11990.43	7414.43	9188.915	9161.97	756.2232	0
5/2/19	9	55	5/2/19	11123.56	B	125833	38.01	2.572	214.6555	6.415	0	2019	2	5	29166.26	4791.74	8607.05	7571.6	3874.176	0
5/2/19	9	52	5/2/19	1150.25	B	125833	38.01	2.572	214.6555	6.415	0	2019	2	5	3490.13	722.87	1672.207	1617.34	428.654	0
5/2/19	9	28	5/2/19	356.9	B	125833	38.01	2.572	214.6555	6.415	0	2019	2	5	600.4	67.59	246.3457	236.3	102.989	0
5/2/19	9	29	5/2/19	2604.7	B	125833	38.01	2.572	214.6555	6.415	0	2019	2	5	5577.07	1001.74	1919.389	1814.04	589.4533	0
5/2/19	9	30	5/2/19	2281	B	125833	38.01	2.572	214.6555	6.415	0	2019	2	5	2469.68	769	1601.398	1623.42	349.8512	0

Store	Dept	Date	Weekly_Sales	Temperature	Fuel_Price	CPI	Unemployment	IsHoliday	Year	Month	Week	max	min	mean	median	std
1	1	5/2/19	24924.5	42.31	2.572	211.096	8.106	0	2019	2	5	57592.1	14537.4	22513.3	18535.5	9854.35
1	11	5/2/19	24213.18	42.31	2.572	211.096	8.106	0	2019	2	5	44553.5	16107.9	24919.3	23607.7	6135.18
1	20	5/2/19	5034.1	42.31	2.572	211.096	8.106	0	2019	2	5	7272.2	2464.49	4091.57	3985.35	921.312
1	4	5/2/19	39954.04	42.31	2.572	211.096	8.106	0	2019	2	5	47893.2	32497.4	36964.2	36580	2930.7
1	8	5/2/19	40129.01	42.31	2.572	211.096	8.106	0	2019	2	5	42663.8	31061.2	35718.3	35356.1	2490.77
1	10	5/2/19	30721.5	42.31	2.572	211.096	8.106	0	2019	2	5	43718.1	23058.4	31033.4	30888.7	3509.19
1	21	5/2/19	8907.63	42.31	2.572	211.096	8.106	0	2019	2	5	13552	5898.29	7808.45	7662.31	1067.25
1	18	5/2/19	4729.5	42.31	2.572	211.096	8.106	0	2019	2	5	53845.1	-1.27	7765.3	2303.36	11435.7
1	2	5/2/19	50605.27	42.31	2.572	211.096	8.106	0	2019	2	5	65615.4	35819.8	46102.1	45561.9	3440.67

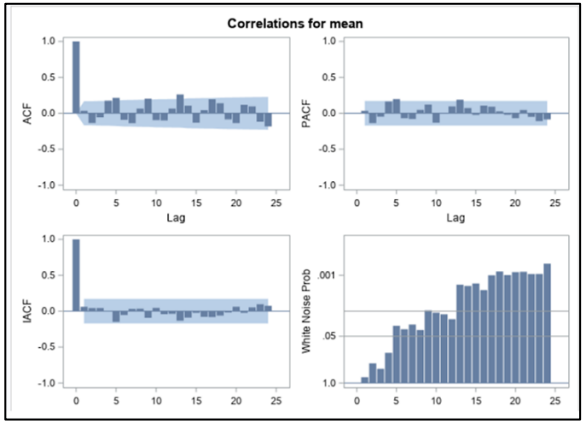
Change the format of the “Date” from dd/mm/yyyy to mm/dd/yyyy.
Removed 4 columns.
We will use only the “1” segment in the “Store” column.

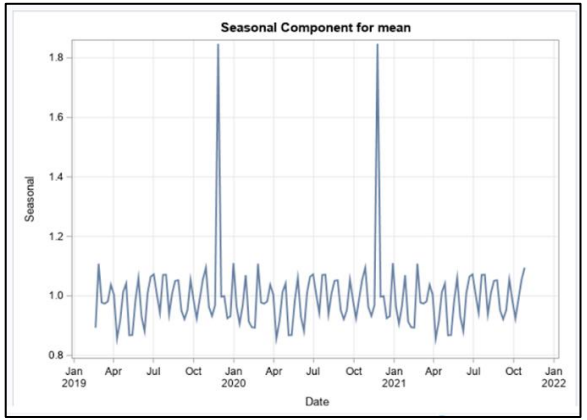
Use dependent variable – “mean”. Independent variables will be “Temperature”, “Fuel_Price”, “CPI”, “Unemployment”, “IsHoliday”.

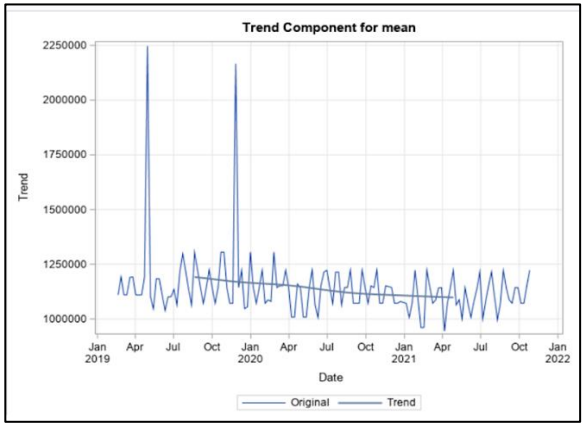
The total variables are 17 and 9000+ rows.

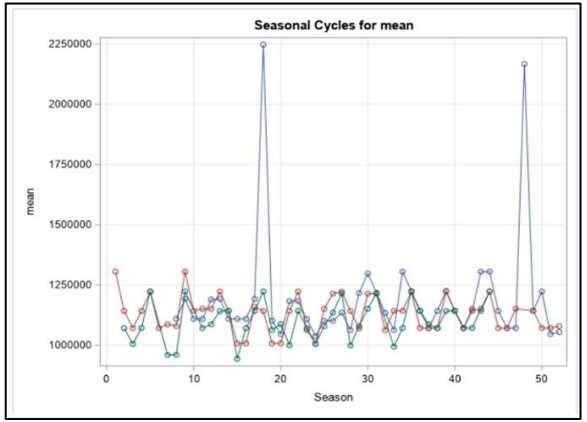
Exploration of Data - Dependent Variable

Dependent variable – “mean”. Accumulation “Sum”. Additional Role – “Date” – “Week”



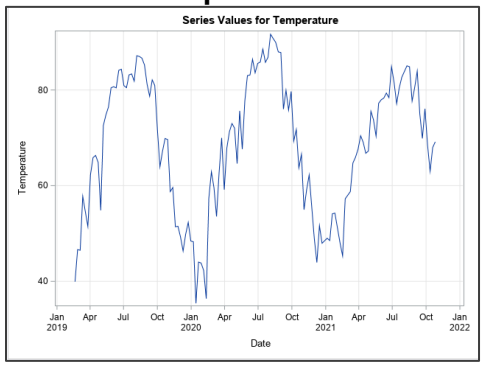




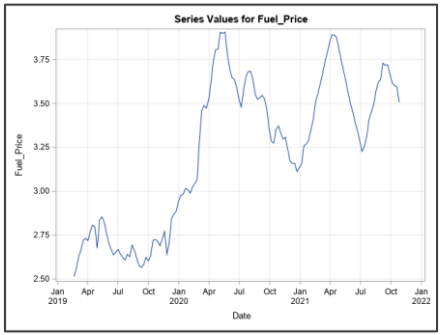


Exploration of Data - Independent Variables

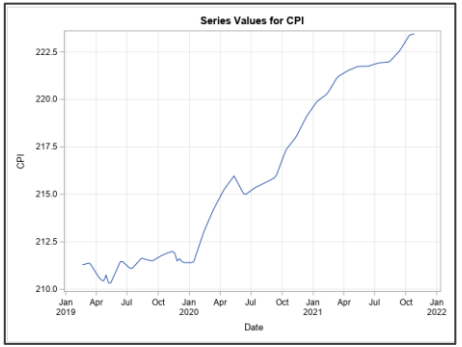
Temperature



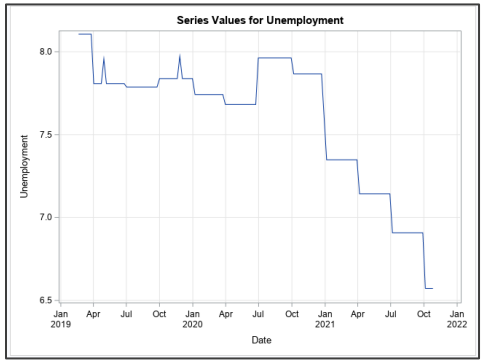
Fuel_Price



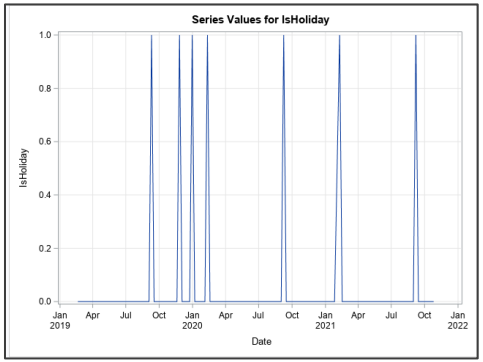
CPI



Unemployment

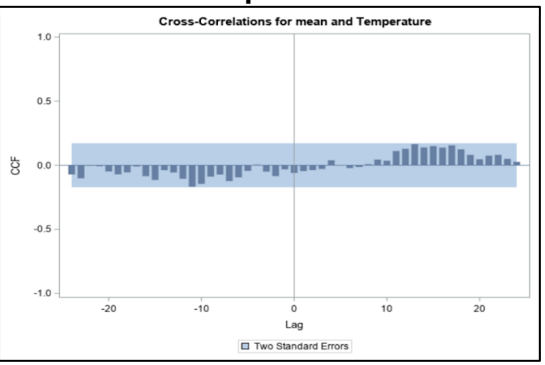


IsHoliday

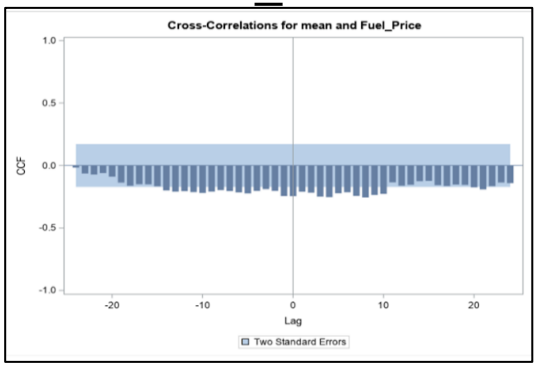


Exploring Variables (Cross-Correlation)

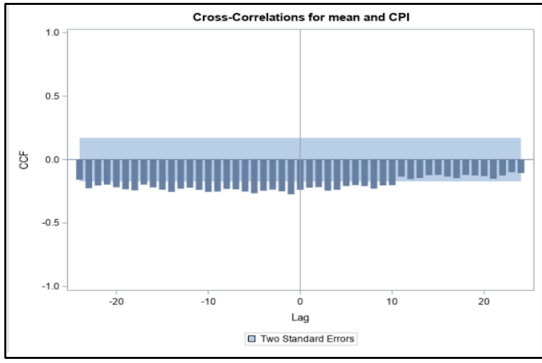
Temperature



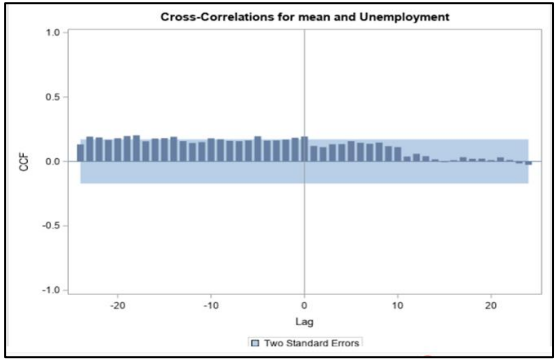
Fuel_Price



CPI

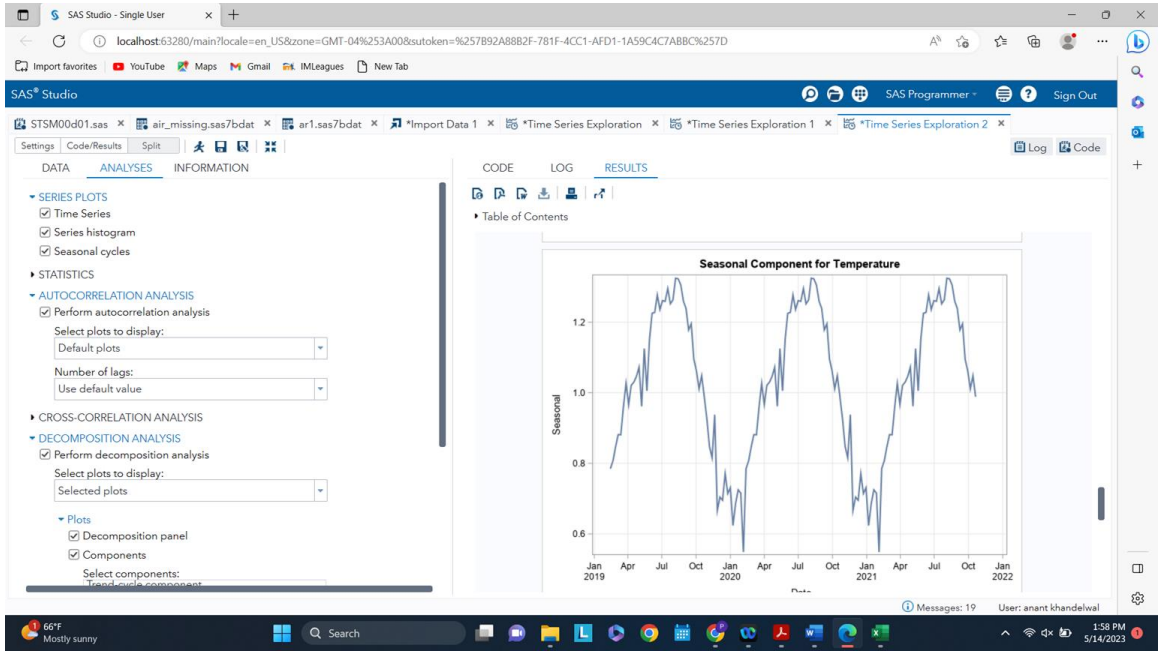


Unemployment



Exploring Variables (Time Series?)

Example Seasonal Component of Temperature



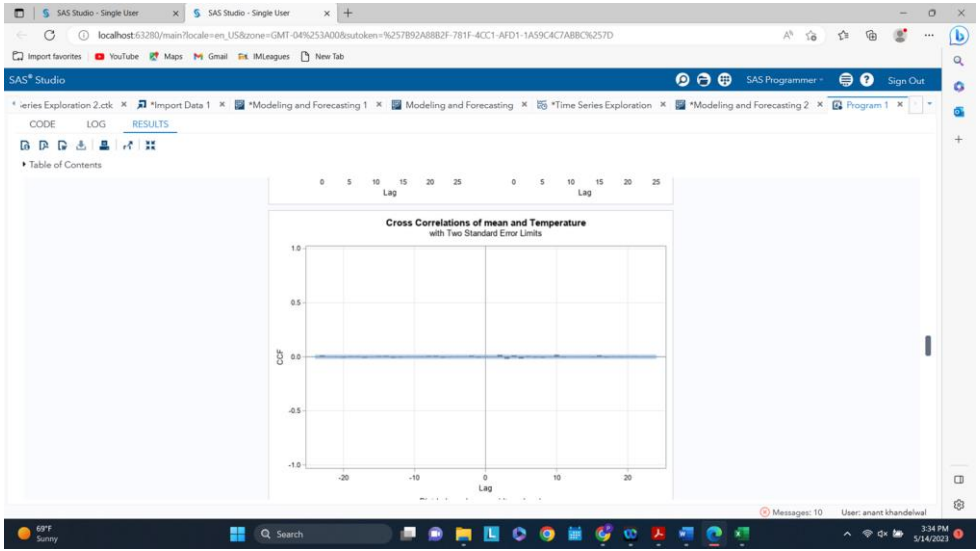
- ➔ Temperature, Fuel Price, CPI, and Unemployment were found to be time series
- ➔ IsHoliday is a categorical variable that is also retained for modeling purposes

Decision

Pre-Whitening Analysis of 4 independent variables

Pre-Whitening

Example Pre-Whitening Cross Correlation between Dependent Variable (Weekly Sales) and Temperature



➔ No Cross Correlation was found for any of the 4 variables

Decision

Will keep or drop independent variables based on the fit metrics

Variable selection for models

Stochastic variables:

Temperature

+

CPI

+

Unemployment

+

Fuel price

Deterministic variables:

Is Holiday

Modeling and Forecasting

Comparing the best set of models to predict our dependent variable:
Mean Weekly sales

1. • ARIMA -> p = 0; q = 2
2. • ARIMAX -> p = 0; q = 0
3. • ARIMAX -> p = 1; q = 0
4. • ARIMAX -> p = 0; q = 1
5. • ARIMAX -> p = 1; q = 1
6. • ARIMAX -> p = 2; q = 0
7. • ARIMAX -> p = 0; q = 2
8. • ARIMAX -> p = 2; q = 2

Pre-whitened variables

+

Exogenous variables

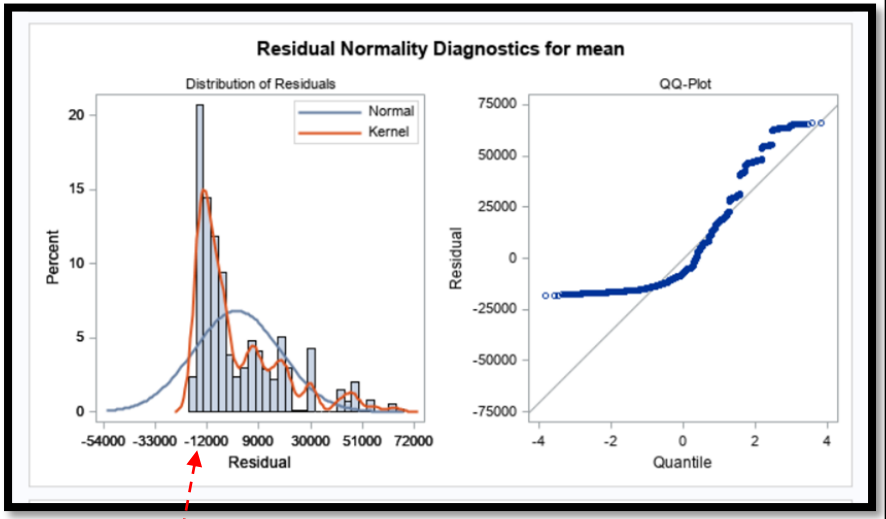
Deep Diving into the accuracy and fit measures

Selecting the lowest MAPE model along with checking the goodness of fit

Models	SSE	n	p	MSE	MAPE	AIC	SBC
ARIMAX(2,0)	3790941127794	9280	8	408859052	2.22%	210334.4	210391.5
ARIMAX(1,0)	4244368791590	9280	7	457712584	4.02%	211380.5	211430.4
ARIMAX(2,2)	2883675746369	9280	10	311076132	4.14%	207808.5	207879.8
ARIMAX(0,2)	2861874675199	9280	8	308657752	4.30%	207734.3	207791.4
ARIMAX(0,0)	2861335610903	9280	6	308533061	4.58%	207740.8	207783.6
ARIMAX(1,1)	2862478387514	9280	8	308722863	4.69%	207736.2	207793.3
ARIMA(0,2)	2862062283406	9280	3	308511618	5.00%	207724.9	207746.3
ARIMAX(0,1)	2862621818212	9280	7	308705038	5.23%	207734.7	207784.7

Checking residuals and white noise

ARIMAX
p = 2; q = 0



Appears right skewed

NO significant autocorrelation exists

Slight white noise

Key conclusions

Optimal model -> ARIMAX(2,0)

MAPE: 2.22%

AIC: 210334.4

SBC: 210391.5

Forecast:

❑ Shows a drop in sales in the forward time horizon following the trend depicted in fit data

Business Recommendations

Given that the macroeconomic indicators show a positive trend, the forecast shows a decline in sales and hence **factors leading to drop** should be investigated

Department level sales trends can be understood to find critical focus areas

Other indicators like quality, delivery of service, price points can be evaluated to understand if there are other factors influencing the sales

Thank You!
Any Questions?