# Model Title:THE ARIMA MODEL APPROACH TO VISUALISE AND FORECAST SALES

[1] Adwin Thomas
*2348404*
*I MDS B*
adwin.thomas@msds.christuniversity.in

[2] Angeline A
*2348409*
*I MDS B*
angeline.a@msds.christuniversity.in

[3] Athul S Kumar
*2348414*
*I MDS B*
athul.kumar@msds.christuniversity.in

[4] Delna M Joseph
*2348419*
*I MDS B*
delna.joseph@msds.christuniversity.in

[5] Gopika Bhaskar ck
*2348424*
*I MDS B*
gopika.bhaskar@msds.christuniversity.in

[6] Jesin K Joy
*2348429*
*I MDS B*
jesin.joy@msds.christuniversity.in

***Abstract*—In this paper, a four-year dataset of a retail global superstore is considered. The plotting of furniture sales against time forms a time series and hence, time series analysis is performed. Forecasting of future values is done with the help of a model called ARIMA or Autoregressive Integrated Moving Average. The parameters that best fit the data are evaluated and fitting is done. Visualisation of the forecast is done and the accuracy of the model obtained can be seen with the help of goodness of fit tests like MSE (Mean Squared Error), RMSE (Root Mean Square Error), and MAPE (Mean Absolute Percentage Error). Hence, the basics of time series analysis are discussed and significance is shown.**

## I. Introduction

In data analysis, a time series is a collection of data points organized in time, where the data points should be equally spaced.
Time-series analysis is a technique used to analyse time-series data and extract meaningful statistical information and characteristics from the data. It is chiefly concerned with identifying three different aspects of the time series, which can be used to better clean, understand, and forecast the data. To do so, it may use a range of models which can process t 2 he time series. One of the major objectives of this is to forecast future values. Extrapolation is involved in the forecasting process when the time series analysis is complex. The forecasted values, alongside the estimation of the associated uncertainty, can make the result extremely valuable.

## II. Benefits of Time Series Analysis

Time series analysis has various benefits for the data analyst. The application of various time series models helps clean, understand data, and forecast future data points.

### A. *Cleaning Data*

The first benefit of time series analysis is that it can help to clean data. This makes it possible to find the true "signal" in a data set, by filtering out the noise. This can mean removing outliers or applying various averages to gain an overall perspective of the meaning of the data. Cleaning data is a prominent part of almost any kind of data analysis. The true benefit of time series analysis is that it is accomplished with little extra effort.

### B. *Understanding Data*

Time series analysis can help an analyst to better understand a data set. This is because the models used in time series analysis help interpret the true meaning of the data, as touched on previously

### C. *Forecasting Data*

A major benefit of time series analysis is that it can be the basis to forecast data. This is because time series analysis — by its very nature — uncovers patterns in data, which can then be used to predict future data points. 3 For example, autocorrelation patterns and seasonality measures can be used to predict when a certain data point can be expected. Further, stationarity measures can be used to estimate what the value of that data point will be. It's the forecasting aspect of time series analysis that makes it so popular in business applications. In addition to analysing and understanding past data, it's being able to predict the future that helps to make optimal business decisions.

## III. Models for Time Series Analysis

Several models can be used to describe and predict data points in a time series. Two of the most basic models used are moving averages and exponential smoothing

### A. *Moving Averages*

A moving average model suggests that an upcoming data point will be equal to the average of past data points. This rudimentary model is powerful in smoothing out data sets to observe their overall trend, with little regard for outlying data points. However, it may smooth out the seasonality of some time series.

### *B. Exponential Smoothing*

Exponential smoothing is a model that predicts upcoming data points based on an exponentially decreasing average of past data points. The Autoregressive and Moving Average (ARMA) model is an important method for studying time series. The Autoregressive Integrated Moving Average (ARIMA) model is based on the ARMA model but converts non-stationary data to stationary data before working on it. ARIMA is widely used to predict linear time series data and offers flexibility in univariate time series model identification, parameter estimation, and forecasting. Time Series analysis is primarily used by businesses and organizations to visualize, understand, and predict losses and gains, making plans accordingly. In this paper, ARIMA modelling is performed to forecast future sales using a dataset of 9800 observations from January 2015 to December 2018.

## IV. Literature Review

### *A. ARIMA Model Building and the Time Series Analysis Approach to Forecasting*

Paul Newbold

Paul Newbold's paper ”ARIMA Model Building and the use of Time Series Analysis Approach to Forecasting” discusses the ARIMA model-building approach to forecasting, focusing on the importance of time series analysis in forecasting. The paper highlights that time series analysis is not an alternative to building regression models for forecasting but rather an integrated approach. The concept of parsimony, which suggests seeking simple structures that describe the major characteristics of the data, is crucial in time series model building.

Box and Jenkins (1970) provided a methodology for fitting time series data as part of ARIMA models. However, there are restrictions on the usage of ARIMA models, such as not considering stationary or stationarity-exhibiting time series after differencing. Furthermore, future predictions are constrained to be linear functions of the observations, as the models are linear.

Despite these limitations, ARIMA models allow for the representation of a wide array of potentially useful predictor functions in models with relatively few parameters, making them parsimonious. However, the need to estimate a large number of parameters may lead to inefficient forecasts.

The initial model selection stage in ARIMA model-building methodology is the most difficult, as statistics calculated often do not provide adequate information. Many dislike the necessity of judgment at this stage and prefer a deterministic scheme yielding a single solution. The paper then discusses the examination of residual autocorrelations from the initially selected model to check the adequacy of representation of a fitted model to the given data.

In conclusion, the paper discusses the usage of ARIMA modelling in time series, the difficulties faced, and recent developments related to other forecasting methodologies.

### *B. Electricity Price Forecasting – ARIMA Model Approach*

Tina Jakasa, Ivan Androcec, Petar Sprcic

This paper on “Electricity Price Forecasting – ARIMA Model Approach” by Tina Jakasa, Ivan Androcec, and Petar Sprcic presents a forecasting technique to model day-ahead spot electricity price using the well-known ARIMA model to analyse and forecast time series. The model is then applied to the time series consisting of day-ahead electricity prices from EPEX power exchange. The day-ahead electricity prices are forecasted using European Energy Exchange data as the reference power market. The hypotheses here test that ARIMA models are good enough for the required forecasting. The original dataset here has 3836 observations over 10 years, against the normally simple and smaller observations used in ARIMA modelling. Hence, this paper attempts to show that such models can be used for data collected over a large period too.

This time series, like many macroeconomic ones, is non-stationary. Therefore, stationarity is obtained using a logarithmic transformation. For the modelling, the Box and Jenkins ARIMA method is used. The paper uses the Expert Modeler in the SPSS software tool which automatically finds the best-fitting model for each dependent series. If independent predictor 8 variables are specified, the Expert Modeler selects and includes only those that have a statistically significant relationship with the dependent series. Automatic detection of outliers is also specified. The Expert Modeler considers seasonal models and this option is only enabled if a periodicity has been defined for the active dataset. The Expert Modeler also includes a constant in the model.

In the end, the best-fitted ARIMA model is (3,0,3) (1,1,1). 3 days are needed to predict the next day's price. 57 outliers were identified and modelled, mainly of the additive (affects a single observation) and transient (impact decays exponentially to 0) types. A review of the outliers identified events that were festivals. MAPE for the model is 3.55Absolute Percentage Error) is 33.1

### *C. Forecasting Crime using the ARIMA Model*

Peng Chen, Hongyong Yuan, Xueming Shu

In this paper on “Forecasting Crime using the ARIMA Model” by Peng Chen, Hongyong Yuan, and Xueming Shu, the ARIMA model was used to make short-term forecasting of property crime for a city in China. The data spans over 50 weeks and the crime amount one week ahead was predicted. The model's fitting and forecasting results were compared with the SES and HES and it was shown that the ARIMA model had higher fitting and forecasting accuracy than exponential smoothing. Such modelling is used by local police stations and municipal governments in decision-making and crime suppression. When this paper was published in 2008, the usage of time series models for short-term forecasting of crime was a new research field and was relatively rare in the world. SES stands for simple exponential smoothing and HES stands for holt two-parameter exponential smoothing.

Not to go too much into details but exponential smoothing is a tool that is often used to fit the trend of the series and it's usually sorted by two types – SES and HES. The statistical 9 software package SPSS is used to formulate the ARIMA model here. The coefficients of the model are solved with the method of conditional least-square estimation.

The paper concludes that the ARIMA model indeed fits the series better than SES and HES. The SES and HES may have depicted the trend of the series well but failed to reflect the fluctuations of the series. The RMSE and MAPE results, too, proved the same. The ARIMA model also approximated the real observation closest and thus proved to be better in terms of forecasting too.

### *D. Forecasting of Indian Stock Market using Time Series ARIMA Model*

Debadrita Banerjee

In this paper on "Forecasting of Indian Stock Market using Time Series ARIMA Model" by Debadrita Banerjee, the ARIMA model was used to predict future stock indices. To establish the model, the validation technique was used on the observed data of Sensex in 2013. The statistical computations and graphical presentation were done with the help of the statistical software SPSS version 15.

The analysis here involved monthly data on the closing stock indices of Sensex over six consecutive years (2007 – 2012) based on which the ARIMA model was formulated to forecast future unobserved indices of Sensex. After obtaining the required data, the Durbin-Watson Test was performed to check whether it was suitable for the required purpose. The Durbin Watson (DW) statistic is a test for autocorrelation in the residuals from a statistical regression analysis.

The Durbin-Watson statistic will always have a value between 0 and 4. If the value lies between 1.5 and 2.5, it is concluded that the data is cross-sectional (independent of time) and regression analysis is to be carried out. If lesser than 1.5 or greater than 2.5, the data is said to be longitudinal (time-dependent) and hence time series analysis is to be applied. In this study, the DW value was 0.121 and thus the data formed a time series with a highly positive 10 correlation.

ARIMA modelling was then carried out with the ARIMA model (1,0,1) and forecasting was done. The limitations of the ARIMA (1,0,1) are also discussed. Firstly, sudden political turbulence or any kind of drastic change in Government policies like fiscal, monetary, or expert input policy will result in higher fluctuation in the Sensex and thus this model wouldn't be able to capture the effect of economic variables for forecasting. Secondly, the data set was assumed to be linear but in reality, that may not be the case. Also, interval forecasting of Sensex was obtained instead of point forecasting.

### *E. RESEARCH ON COVID 19 EPIDEMIC BASED ON ARIMA MODEL*

Li Zhihao et al 2021 J. Phys.: Conf. Ser. 2012

The COVID-19 pandemic has caused significant global impacts, including the spread of fever, dry cough, fatigue, and dyspnea. Despite efforts to control the spread, uneven development levels have exposed shortcomings in prevention and control. Countries like China and the United States have developed anti-coronavirus vaccines, but the situation remains grim. Current predictions of the new crown epidemic are based on infectious disease transmission dynamics models, such as SEIR and SIR models, and statistical models like time series analysis.

The autoregressive integrated moving average model (ARIMA) is a widely used time series model with strong short-term predictability and simplicity. It is built by transforming a non-stationary time series into a stationary one, redressing the lag value of the series and the present value and lag value of the random error term. The ARIMA model has high prediction accuracy of short-term trends, making it useful in epidemic trend prediction of infectious and non-communicable diseases.

The ARIMA model was used to establish prediction models of new cases, cure rate prediction, and death in the United States from April 3 to April 12, 2021. The results showed that the ARIMA (1, 1, 2) optimal model was closer to the actual observed value for the number of new cases, with an average error of 7056.2. The ARIMA model is more accurate for forecasting results and simpler than traditional SEIR models. However, it requires too high residual white noise testing for appropriate data, resulting in fewer predictable indicators.

### *F. Forecasting of demand using ARIMA model*

Jamal Fattah and team

In today's competitive manufacturing environment, organizations are focusing on demand-driven supply chains to respond quickly to shifting demand. Demand forecasting is crucial for inventory management, as inaccurate estimations can lead to significant costs and stock outs. Intermittent demands present challenges for traditional statistical demand forecasting methods, as they require historical data.

An ARIMA model was developed to model demand forecasting in food manufacturing using the Box-Jenkins time series approach. The model, which minimizes the four performance criteria, is ARIMA (1, 0). The results show that this model can be used for modeling and forecasting future demand in food manufacturing, providing reliable guidelines for managers in making decisions.

Future work will involve developing other models using a combination of qualitative and quantitative techniques to generate reliable forecasts and increase forecast accuracy. The neural network approach will also be compared with ARIMA's results to confirm its strength in the food company. An ARIMA-radial basis will be created to achieve high accuracy.

In conclusion, demand forecasting is an essential function in managing supply chains and is crucial for businesses to deploy for the future. Future work will focus on developing models using a combination of qualitative and quantitative techniques and neural network approaches to achieve high accuracy in forecasting.

### *G. Time Series forecasting using ARIMA model: Case study of Mining face Drilling rig*

Hussan Al-Chalabi, Yamur K. Al-Douri and Jan Lundberg

Time series forecasting is a method that predicts future data points based on observed data over a period known as lead-time. It is used to provide a basis for production control, production planning, and optimization of industrial processes and economic planning. Traditional models, such as the Box-Jenkins or the Autoregressive Integrated Moving Average (ARIMA) model, assume time series data are generated by linear processes. However, these models may be inappropriate if the underlying mechanism is nonlinear.

A hybrid method combining ARIMA and Artificial Neural Network (ANN) models has been proposed to improve forecasting accuracy. This method outperforms each component model, and both require a large sample size to build a successful model.

An ARIMA model was used to forecast the Total Cost (TC) data for a face drilling rig used in an underground mine in Sweden. Findings from the case study suggest that the ARIMA model is appropriate, but the parameters need to be better estimated for accurate forecasting.

The ARIMA parameters (p, d, q) were used in four different scenarios, and the parameter values had a strong effect on the forecasting method. Therefore, these parameters need better estimation from the data for accurate forecasting. AI such as MOGA based on the ARIMA model could provide other possibilities for estimating the parameters and improve data forecasting. The MOGA based on the ARIMA model can be used to forecast data with high accuracy and can be used for life cycle cost analysis.

## V. Modelling

### *A. ARIMA Models*

The Autoregressive Integrated Moving Average (ARIMA) model is a linear model that uses time-series data and statistical analysis to interpret and make future predictions. It aims to explain data by using time series data on its past values and using linear regression to make predictions. The key components of the ARIMA model are the ”AR” stands for autoregression, which indicates that the model uses the dependent relationship between current data and its past values.

The ”I” stands for integrated, meaning that the data is stationary. Stationary data refers to time-series data that's been made “stationary” by subtracting the observations from the previous values. The “MA” stands for moving average model, indicating that the forecast or outcome of the model depends linearly on past values.

Each of the AR, I, and MA components are included in the model as a parameter, assigned specific integer values that indicate the type of ARIMA model. A common notation for the ARIMA parameters is: ARIMA (p, d, q).

The formulating of the ARIMA model is a complicated process, but in summary, it includes four steps: (1) Identification of the ARIMA (p, d, q) structure.
The Akaike information criterion (AIC) is a mathematical method for evaluating how well a model fits the data it was generated from. In statistics, AIC is used to compare different possible models and determine which one is the best fit for the data. AIC helps select the model which best explains the variance in the dependent variable with the fewest possible independent variables.

AIC helps reduce overfitting by reducing the number of parameters used to build the model and the maximum likelihood estimates of the model. It helps reduce the cost of adding any given parameter and measures the information lost, so the model with a lower AIC score indicates a better fit.

The ARIMA model could provide forecasting results with upper limits, lower limits, and forecasted values. The upper and lower limits provide a confidence interval of 0 to 1 and a probability of 0.5.

In conclusion, the ARIMA model is a linear model that uses time-series data and statistical analysis to interpret and make predictions. It is based on the Autoregressive Integrated Moving Average (ARIMA) model, which is a linear model that is suitable for dealing with stochastic series. The model can be further improved by using the Akaike Information Criterion (AIC) values, which help determine the best-fit model for the data.

### B. *Box-Jenkins Model Identification*

The Box-Jenkins model is a statistical method used to analyze time series data. It involves determining whether the time series is stationary and if there is any significant seasonality that needs to be modelled. Stationarity can be assessed using run sequence plots or autocorrelation plots, while seasonality can be assessed using autocorrelation, seasonal subseries plots, or spectral plots.

To achieve stationarity, Box and Jenkins recommend the differencing approach, but fitting a curve and subtracting the fitted values from the original data can also be used. At the model identification stage, the goal is to detect seasonality and identify the order for the seasonal autoregressive and moving average terms. For Box-Jenkins models, seasonality is not explicitly removed before fitting the model, but the order of the seasonal terms is included in the model specification to the ARIMA estimation software.

Seasonal differencing is used for seasonal differencing of a time series, similar to regular differential, but instead of subtracting consecutive terms, the value from the previous season is subtracted. The model is represented as SARIMA (p, d, q) x (P, D, Q), where P, D, and Q are the SAR (seasonal autoregressive), order of seasonal differencing, and SMA (seasonal moving average) terms, respectively. If the model has well-defined seasonal patterns, D=1 is enforced for a given frequency ’x’.

### C. *Measures for Goodness of Fit*

To evaluate the prediction performance, it is necessary to introduce a forecasting evaluation criterion. In this study, quantitative evaluation is used as the accuracy measure.

*1) Root Mean Square Error(RMSE):*

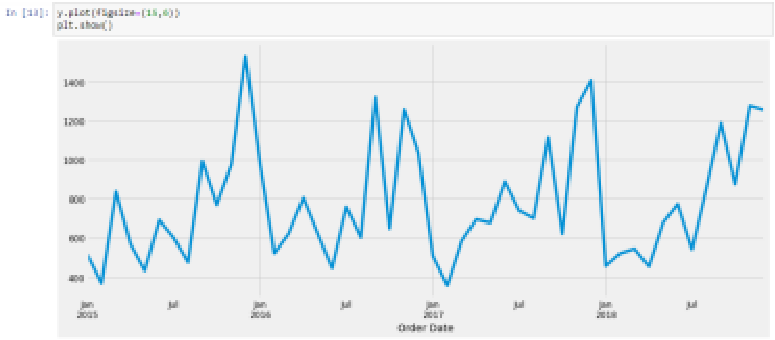
$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \tag{1}$$

*2) Mean Absolute Percentage Error (MAPE):*

$$\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}_i}{y_i}\right| \times 100\% \tag{2}$$

*3) Mean Absolute Error (MAE):*

$$\text{MAE}=\frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i| \tag{3}$$



```
In [13]: y.plot(figsize=(15,6))
         plt.show()
```

Fig. 1. Upward Trend

*4) Ljung-Box Q Statistic:* Overall adequacy is provided by the Ljung-Box Q statistic.

$$Q=n(n+2)\sum_{k=1}^{h}\frac{\hat{\rho}_k^2}{n-k} \tag{4}$$

Where, $r_k(e)$ = the residual autocorrelation at lag k 18 n = the number of residuals m = the number of time lags included in the test. If the p-value associated with the Q statistics is greater than 0.05, then the model is considered to be adequate.

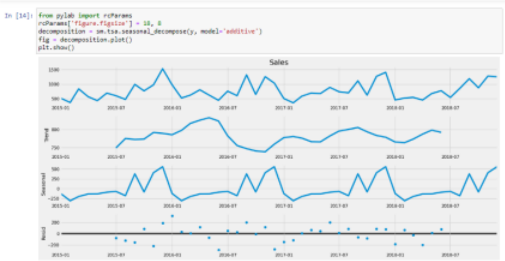
### D. *Discussions*

As mentioned above, a four-year dataset of a retail superstore is taken on which an ARIMA model is fitted and one-step ahead forecasting is done. The idea of setting up a one-step-ahead forecast is to evaluate how well a model would have done if you were forecasting for one day ahead, during 5 years, using the latest observations to make your forecast. Forecasting and measure of goodness of fit are performed on Furniture Sales Time Series Data. The following are the results obtained.

*1) Visualizing Furniture Sales Time Series Data:* The data plotting shows distinct patterns, such as seasonality, with sales always low at the start and high at the end of the year. An upward trend is present within a single year, with some low months in the middle. Visualizing this can be done by decomposing the time series into its components(Refer Fig.1). Here, the time series is additive and is decomposed into its components – Trend, Seasonality, and Noise (irregular variation). The sales of furniture are unstable due to a large number of irregular variations and its seasonality is very obvious. There was an overall increasing trend between the mid of 2015 to mid of 2016 which went down sharply in late 2016 after which a gradual increase in trend can be seen.(Refer Fig.2)

*2) Forecasting and Validation of Forecasts:* To understand the accuracy of the forecast, predicted sales are compared to real sales of the time series. The forecasting is set to start on 2018-01-01 to the end of the data.

The line plot shows the observed values compared to the



```
In [14]: from pylab import rcParams
         rcParams['figure.figsize'] = 18, 8
         decomposition = sm.tsa.seasonal_decompose(y, model='additive')
         fig = decomposition.plot()
         plt.show()
```

Fig. 2. Trend,Seasonality and Noise

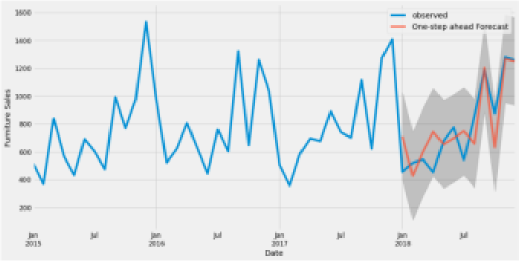


Fig. 3. Line plot

rolling forecast predictions. Overall, the predictions align quite well with the true values. An upward trend is still captured in the forecast, from the beginning of the year and the seasonality is obvious too. The grey region depicts the confidence interval of the forecast.(Refer Fig.3)

### E. Conclusion

In this paper, a time series analysis of furniture sales over four years was performed. ARIMA modelling was done and the best-fitted model turned out to be SARIMAX (1,1,0,12) which satisfied all the tests under the goodness of fit. Forecasting was done using the model which turned out to be pretty accurate with the same seasonality as that present in the original dataset. Hence, I learned the basics of ARIMA modelling and time series forecasting, which form the basis of data science.

Time series analysis and forecasting are extremely crucial for the functioning of various organizations and processes in the world and the ARIMA model is one such model used. However, ARIMA can be limited in forecasting extreme values. While the model is adept at modelling seasonality and trends, outliers are difficult to forecast for ARIMA as they lie outside of the general trend captured by the model. In the end, different models are suited for different time series and no one model can be applied to every time series out there.

### F. References

1. Banerjee, Debadrita. ”Forecasting of Indian Stock Market using Time-series ARIMA Model” - 2014 2nd International Conference on Business and Information Management (ICBIM) (2014)

2. Bevans, Rebecca. Akaike Information Criterion – scribbr (25th May 2022).

3. Box, George EP, and George C. Tiao. ”Intervention analysis with applications to economic and environmental problems.” Journal of the American Statistical Association 70.349 (1975): 70-79

4. Bush, Thomas. ”Time Series Analysis: Definition, Benefits, Models” - pestle analysis (8th June 2008)

5. CFI Team. ”Autoregressive Integrated Moving Average (ARIMA)” - Corporate Finance Institute (3rd November 2022)

6. Chen, S., et al. ”The time series forecasting: from the aspect of network.” arXiv preprint arXiv:1403.1713 (2014)

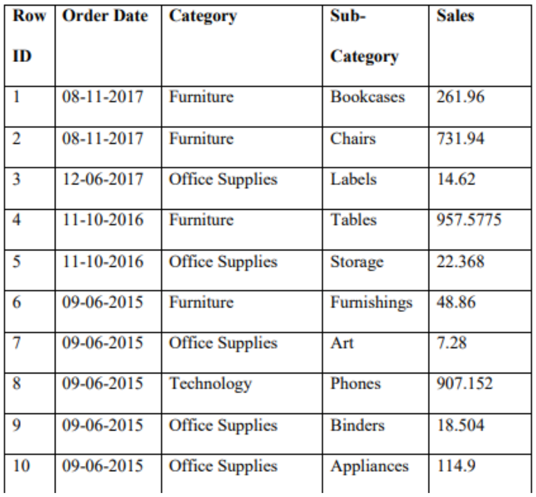
7. George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, Greta M. Ljung - Time Series Analysis: Forecasting and Control (Wiley Series in Probability and Statistics) 5th Edition

8. Goswami, Saptani. “Study of Effectiveness of Time Series Modelling (Arima) in Forecasting Stock Prices” – Research Gate (2014)

9. Jakasa, Tina; Androcec, Ivan; Sprcic, Petar. ”Electricity price forecasting - ARIMA model approach” - 2011 8th International Conference on the European Energy Market (EEM) (2011)

### G. DATASET

The first 10 data of dataset,Refer Fig.4



| Row ID | Order Date | Category | Sub-Category | Sales |
|---|---|---|---|---|
| 1 | 08-11-2017 | Furniture | Bookcases | 261.96 |
| 2 | 08-11-2017 | Furniture | Chairs | 731.94 |
| 3 | 12-06-2017 | Office Supplies | Labels | 14.62 |
| 4 | 11-10-2016 | Furniture | Tables | 957.5775 |
| 5 | 11-10-2016 | Office Supplies | Storage | 22.368 |
| 6 | 09-06-2015 | Furniture | Furnishings | 48.86 |
| 7 | 09-06-2015 | Office Supplies | Art | 7.28 |
| 8 | 09-06-2015 | Technology | Phones | 907.152 |
| 9 | 09-06-2015 | Office Supplies | Binders | 18.504 |
| 10 | 09-06-2015 | Office Supplies | Appliances | 114.9 |

Fig. 4. DATASET1