VISUALIZING TIME SERIES DATA

Sathya Krishnan Suresh Shunmugapriya P

satyakrishnan.s@pec.edu pshunmugapriya@gmail.com

**ABSTRACT**

Time series is a collection of data instances that are ordered according to a time stamp. Visualizing the time series data constitutes an important part of the process for working with a time series dataset. Visualizing the data not only helps in the modelling process but it can also be used to identify trends and features that cause those trends. The code for the visualizations used in this article can be found in my GitHub repository – [Visualizing Time Series Data](https://github.com/SathyaKrishnan1211/visualizing-time-series-data) [2] and the dataset used can be found [here.](https://www.kaggle.com/competitions/tabular-playground-series-mar-2022/data)[3]

**INTRODUCTION:**

Visualization in the area of time series is a concept that has not been focussed on a lot and in this article we try to develop a general method that could be used for most of the time series datasets. In this article, the statistical components of a Time Series are not analysed using visualization but using the method described here, the statistical components can also be analysed. Some of the terms mentioned that can be confusing are ‘unit of measurement’ – the measurement of time chosen to analyse like hour, day and ‘splitting the plot’ - dividing the a single line into multiple lines based on a third parameter.

**TIME SERIES**

Time series is a collection of data instances that are ordered according to a time stamp, or it is a measurement of an item over a period. For example, the price of oil for the last three months represents a time series where the price of oil is the ‘item’ in definition. Time Series are widely used in econometrics, weather and price forecasting, etc. It is made up of four components – Trend, Seasonal Variations, Cyclic Variations and Random or Irregular movements.

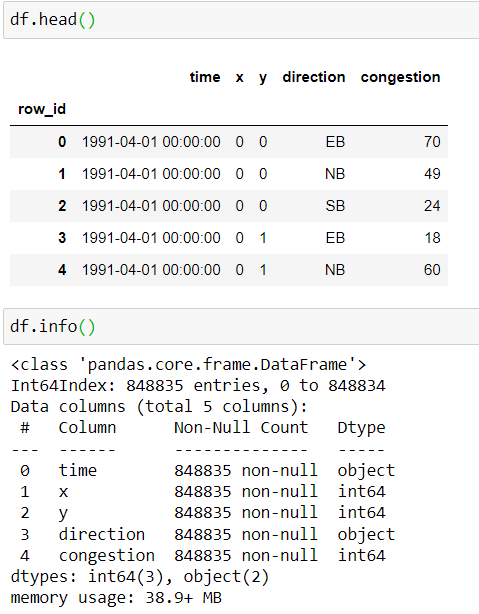
Trend represents the general increase or decrease of the quantity over the period for which the data was collected. Trend can either be upward, downward, or stable.

Seasonal variations represent the variations in the values of the quantity for a particular period and it is likely to happen the next time the period arrives.

Random or irregular movements account for the things that we cannot predict. For example, the occurrence of a natural disaster altering the trend component of the time series.

**DATASET DESCRIPTION:**

The dataset that has been used for writing this article is taken from a Kaggle competition called “Tabular Playground Series – Mar 2022”. The dataset used in the competition has been derived from [Chicago Traffic Tracker – Historical Congestion Estimates dataset](https://data.cityofchicago.org/Transportation/Chicago-Traffic-Tracker-Historical-Congestion-Esti/sxs8-h27x). Figure 1 shows the features of the dataset and the count and datatype of each feature. This is the data before it has been pre-processed.



*Figure-1*

The dataset consists of the measurement of traffic congestion across 65 roadways from April through September of 1991. Here the *row\_id* is a unique identifier for each instance, time is the 20-minute period in which the measurement was taken, x is the east-west midpoint coordinate of the roadway, y is the north-south midpoint coordinate of the roadway, direction is the direction of travel and finally the target variable congestion which represents the congestion levels for the roadway during each hour. The values for congestion have been normalized to range from 0 to 100. The goal of this article is to use the time stamps to derive information about congestion using visualization and in the process of deriving the information develop methods that can be generalized to similar problems.

**DATA PREPARATION:**

Data in its original form cannot be used for visualization and modelling process most of the time. Data should be cleaned, imputed and converted to a form preferred for the problem [4]. As you can see from Figure-1 the data in its original form cannot be used for visualization as the year, month, date and time values have been put together in a string. Those values have to be separated and new columns have to be created for each of the above-mentioned features. First the date attributes like month, date and year have been separated from the ‘time’ attribute and they have been added as separate features as you can see from Figure-2. Next, the time related attributes like hour, minute and second have derived from the ‘time’ attribute and they have been added as features to the dataset.

A screenshot of a computer

Description automatically generated with medium confidence

*Figure-2*

After forming these attributes from ‘time’ feature, we can then extract some more details like weekday or weekend, day of a particular instance and many other details. Extracting as many details as possible without repeating the same information is one of the most important parts of data preparation. The final form of the dataset before using it for data visualization and modelling is shown in Figure-3.

A picture containing graphical user interface

Description automatically generated

*Figure-3*

**TIME SERIES VISUALIZATION:**

Time series visualization is the process of plotting out the variations of the target feature over some period of time which maybe measured in months, days and hours depending on the target feature. For example, stock prices will be reported daily whereas the change in ranking of a tennis player is reported on a weekly basis.

This section of the article will focus on how the combination of features can be used to represent the variation of congestion over a period. First, let’s start with a simple plot. The variation of congestion with respect to month has been shown in Figure-4. We can infer that till June, the traffic flow was following a decreasing trend but it suddenly increased for the month of August indicating that during the month of August there might have been some important event or festival in Chicago.

Chart, line chart

Description automatically generated

*Figure-4*

We can build upon the previous plot by splitting the average congestion based on two features. The first feature is the ‘PM’ in the dataset and the second feature is the ‘Weekday’ column. Figure-5 compares the average congestion between weekdays and weekends and between AM and PM. Figure-5(ii) shows that the average congestion was less during the weekends, and it was higher during the weekdays maybe indicating that in 1991 Chicago always had a busy week. Figure-5(i) shows that as the day progressed the average congestion kept on increasing.

From the above and below figures we can infer very little with ‘month’ as the unit of measurement. Choosing a proper unit of measurement is often very important when it comes to Time Series Analysis.

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

*Figure-5(i & ii)*

Since ‘month’ is not the best unit of measurement, ‘hour’ is used as the next unit of measurement which is shown in Figure-6. Figure-6 and figure-5(i) confirm that congestion increases as the day progresses and it starts to fall off around 6.00 PM. In general, whenever you are trying to measure a quantity like traffic congestion against some unit of time, ‘hour’ is considered as to be the best unit as it lies in the middle of both extremes like days and minutes.

Chart, line chart

Description automatically generated

*Figure-6*

Just like how Figure-5 was developed from Figure-4, the following figure (Figure-7) is developed from the previous figure. In Figure-6 the average congestion with respect to hour has been plotted but this does not show us what is the variation for each day. Figure-7 splits Figure-6 into the seven days of the week and the plot gives us some interesting insights into the congestion statistics of Chicago. From the plot, you can see that all the days start with the same congestion but as the day progresses weekdays have a sudden spike in their congestion values. While Saturday and Sunday also follow a similar pattern their values are far less than those of the weekdays.

Chart, line chart

Description automatically generated

*Figure-7*

Figure-8 uses the direction of travel to split the variation of congestion with respect to month and in this figure there is a distinct separation between the congestion values for each direction. These sort of plots which separate out the target variable distinctly are very important for time series analysis as they enable us to confidently forecast values.

Chart, line chart

Description automatically generated

*Figure-8*

With the possible combinations of the co-ordinate features ‘x’ and ‘y’, a new dataframe is formed with ‘month’ as row and the combinations of ‘x’ and ‘y’ as columns. Features ‘x’ and ‘y’ by themselves won’t contribute much to our understanding of the variation of congestion with respect to co-ordinates and therefore their combinations have been used. Each entry in the dataframe represents the average congestion corresponding to that particular month and that particular co-ordinate. Figure-9 can be considered as an important plot as it can be used to predict which co-ordinates will require more surveillance and which do not, thereby optimizing the allocation of resources. Figure-9 consists of a lot of information, and it can be further decomposed to infer more information about each co-ordinate. Only four of the subplots have been shown here. To view all the eight subplots kindly refer to the source code.

A picture containing window, building, light, overhead

Description automatically generated

*Figure-9*

And finally, Figure-10 plots out variation of congestion for each day of a month. Analysing the trend of a single day will do no good but analysing them as groups (for example, first 10 days of the month, last 10 days of the month, etc) will give us a lot of information. These types of plots require a lot of time to analyse as a lot of information is squeezed into a single plot. So, often these types of plots are decomposed into groups mentioned above and they are analysed separately.

Chart

Description automatically generated

*Figure-10*

**METHOD TO VISUALIZE TIME SERIES DATA:**

In the previous section, a lot of plots were shown and the inferences drawn from them were pointed out. From the previous section, a general method to visualize time series data is provided in this section.

1. First and foremost, thing is to select the unit of measurement like hour, day, month that suits the target variable and gives us a lot of scope for analysis.
2. After selecting the unit of measurement, create a plot with the unit of measurement along x-axis and the target variable along the y-axis.
3. Use the categorical features in the dataset as ‘hue’ parameter to split the plot mentioned in step 2. If the plot becomes crowded split the ‘hue’ parameter into groups and analyse them. The purpose of splitting the plot is to know how the target variable varies during a particular period according to the ‘hue’ parameter.
4. For the continuous valued features, split them into bins using Sturge’s rule,

where N is the number of instances. Then use those bins as ‘hue’ parameter to split the plot. This division of continuous valued columns will also showcase the trends in the target variable.

1. After completing step 3 and 4, pick the features that form the most distinct classes in the plot and try to use the combination of these features as the ‘hue’ parameter to split the plot further.

**CONCLUSION:**

There are a lot of ways to visualize a time series data and in this article, I have tried to provide a method that can be applied to most of the time series datasets.

**REFERENCES:**

1. (2008) Time Series. In: The Concise Encyclopedia of Statistics. Springer, New York, NY. <https://doi.org/10.1007/978-0-387-32833-1_401>
2. Kernel referred: <https://www.kaggle.com/code/sathyakrishnan12/kaggle-tps-march-22>
3. Dataset used: <https://www.kaggle.com/competitions/tabular-playground-series-mar-2022/data>
4. Abdallah Z.S., Du L., Webb G.I. (2017) Data Preparation. In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning and Data Mining. Springer, Boston, MA. <https://doi.org/10.1007/978-1-4899-7687-1_62>
5. Barapatre, D., and V. A. “DATA PREPARATION ON LARGE DATASETS FOR DATA SCIENCE”. *Asian Journal of Pharmaceutical and Clinical Research*, vol. 10, no. 13, Apr. 2017, pp. 485-8, doi:10.22159/ajpcr.2017.v10s1.20526.
6. Yujie Fang et al 2020 IOP Conf. Ser.: Mater. Sci. Eng. **782** 022013
7. W. Aigner, S. Miksch, H. Schumann, and C. Tominski, Visualization of time-oriented data .Springer Science & Business Media, 2011
8. T. Blascheck et al. , "Exploration Strategies for Discovery of Interactivity in Visualizations,"IEEE transactions on visualization and computer graphics, vol. 25, no. 2, pp. 1407-1420,2018.