Manipulation of Delhi Environment Data for Weather Prediction

Adithya Reddy, 15BCE1055
School of Computing Sciences and Engineering, VIT Chennai, Tamilnadu, India,600127
Email:chreddy97@gmail.com

***Abstract*— Data Manipulation is the first step in machine learning. This step itself doesn't provide any usable outputs, but it is very crucial for further analysis that we perform on the data. Data from real life is generally very noisy. It may contain null values, trivial data and false values due to error in recording equipment. We must remove noise from the data and make it clean so that it can be used for learning. In this experiment I have taken weather data of Delhi. This data will be used to predict whether the outcome is haze or fog depending on temperature and humidity values.**

I. Introduction

Machine Learning algorithms for most of the cases are designed to work for organized data and might throw an error if there are null values in data or predicts a function with significant errors. To avoid this, we import data collected from real life, observe the data and make necessary changes both general and specific to the problem and then use the new data for learning data. In the Delhi weather data set, there are a lot of parameters affecting each other and also weather outcome. To make the problem simpler we consider only weather conditions haze and fog and assume that it is mostly represented by temperature and humidity most of the time. Although it is certain that other factors have an impact on weather condition, it is clear from the data that temperature and humidity have the most effect.

II. Methodology

We import the data from csv file and convert it to a data frame that is ready for manipulation. We plot graph of all pairs to check which parameters determine the weather. We take clue from the graphs for which weather conditions to choose for classification. For example, light rain and heavy rain may have almost similar conditions and there is no point in classifying them. Then we remove the rows with null values and we also remove outliers if possible.

III. Database

The database contains different attributes of environment like temperature, pressure, humidity and wind direction. All these parameters determine whether the weather is clear, windy, stormy, foggy etc. To simplify our analysis, we consider only haze and fog as they have very contrasting values in terms of temperature and humidity.

IV. Experiments

Initially we plot graphs between different parameters to observe patterns between the parameters. (Figure.1)

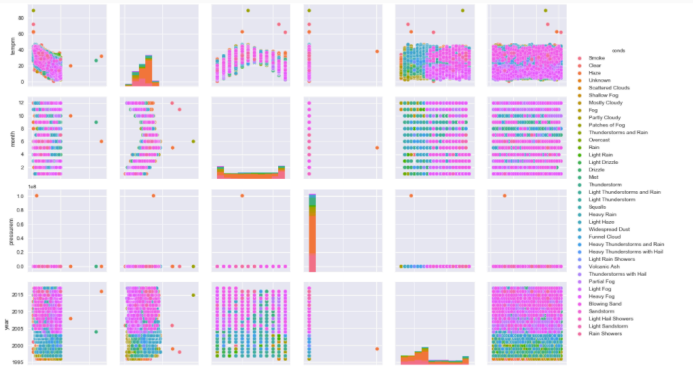

Figure.1

As it is evident from the figure it is too complicated to classify those many weather conditions. So we reduce the complexity of data by selecting a few weather conditions and by elimination null values.(Figure.2)

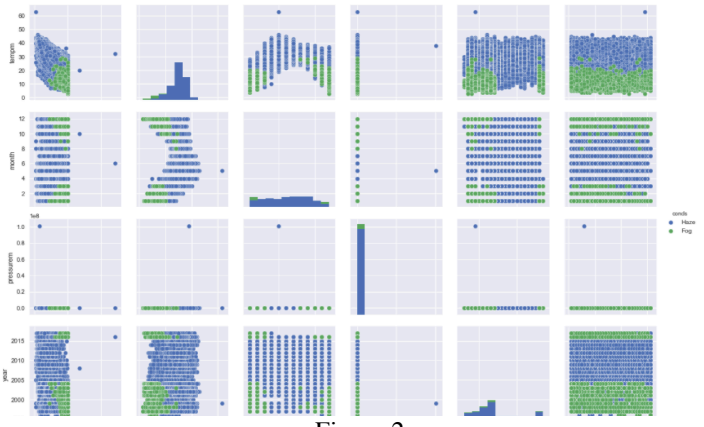

Figure.2

Violin plots are used to represent variable distribution.(Figure.3)

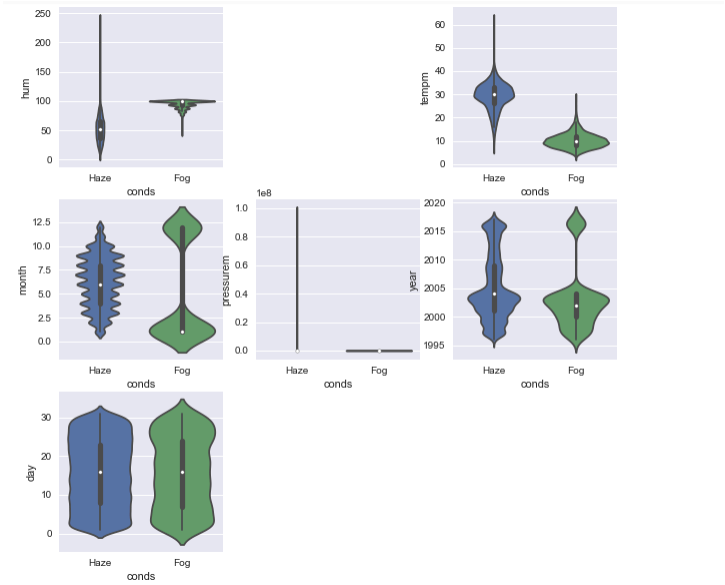

Figure.3

Box plot is another way of representing distribution(Figure.4)

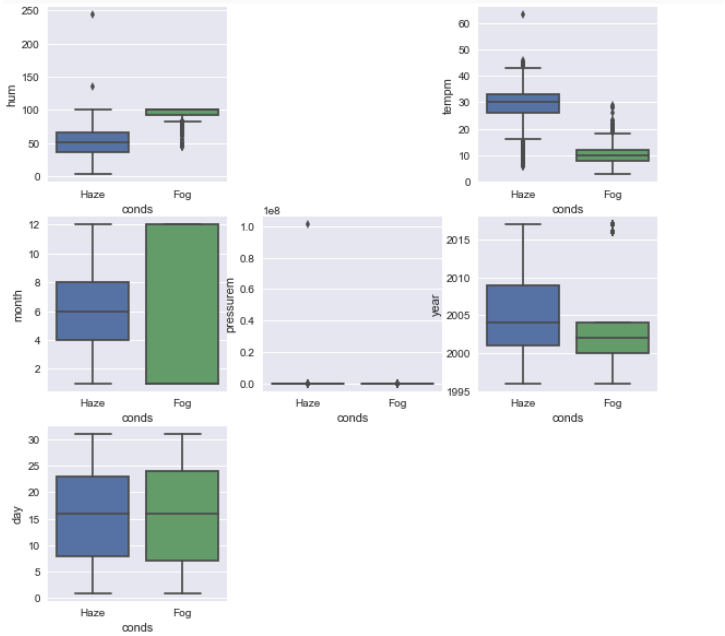

Figure.4

From violin plot or box plot we can observe that humidity and temperature can be used to differentiate between fog and haze but other factors are not as useful.

V. Conclusion

The data is separable to an extent when we take haze and fog into consideration with temperature and humidity as parameters. The data is separable even in case of temperature and days (in a month) as parameters. But we are more interested in finding weather using environment conditions rather than drawing a graph to find which day has which weather. Other parameters are not considered as make the data inseparable.

VI. References

[1] https://www.kaggle.com/mahirkukreja/delhi-weather-data