MULTIMEDIA UNIVERSITY

TDS 3301 DATA MINING

ASSIGNMENT 1

EXPLORATORY DATA ANALYSIS

GROUP DETAILS

|  |  |  |
| --- | --- | --- |
| NAME | ID | EMAIL |
| Darrel Shakri Bin Ahmad Shakri | 1141327906 | menubearer@gmail.com |
|  |  |  |
|  |  |  |
|  |  |  |

ABOUT THIS DATASET

Name: Bike Sharing Dataset

URL of dataset: <https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset>

DESCRIPTION

This dataset contains information on weather information such as windspeed and temperature, and the number of mobile rental bikes. It is recorded every hour, and other time information is also recorded such as season, month, and whether the day is a weekday, or is a holiday.

There are two .csv files in the dataset. One is hour.csv, which records, every hour, the amount of bikes that is on the move, and weather information for that hour. The dataset spans two years, from 2011 to 2012, resulting in a lot of observations. Another dataset, day.csv, aggregates hourly observations belonging to the same day and turn it into a daily dataset, with bike counts summed up.

POSSIBLE INSIGHTS

This dataset represents the mobility of the citizens in the city. As more and more cities adopt bike sharing systems for the public due to health and environment issues, more bikes are being used as commute around the city.

By mining this dataset, we can find out which weather, seasonal, and time of day characteristic leads to higher mobility around the city, represented by the increase in the number of bicycles that are rented and on the move. Further, after creating these predictions, we can detect significant events that defies this prediction, and can be used for validation of said significant event.

BEST MINING METHOD

The best mining method for obtaining the insights mentioned above can be **clustering**. By clustering various weather, time, or season information against the number of bikes in transit, we can more easily identify which of these variables lead to increased mobility in the city, represented by the number of rented bicycles.

DATA QUALITY ISSUES

While the dataset has apparently been preprocessed for use in a research paper, there are minor quality issues. The main issue is with regards to the day.csv dataset, which is an “summed up” dataset from the hour.csv. The assignment of the weather situation variable in the day.csv is unclear. For example, the first day of day.csv, 2011-01-01, has the weathersit variable set to 2, but according to the hour.csv, most of the hours in that day has its weathersit variable set to 1. There is no explanation in the URL stated above why that is so.

The rows in the day.csv dataset recorded all the hours in the hour.csv. If one wants to perform data mining based on day.csv, the time periods where people would be asleep will be not as useful as the time period where people are awake and is commuting back and forth. Therefore, another data transformation may be necessary.

DATASET VISUALS

The visualization can be viewed as a Shiny web application over at the following URL:

<https://mmudsask.shinyapps.io/bikeshare/>

There are various visuals, such as weather situation over time, and you can filter the dataset by year and season.