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
Nur Nadhirah Bt. Nazarudin	1142700151	nazanadhirah@gmail.com
Andy Yong Jun Jie	1142701565	Andyyong9611@gmail.com
Kirbashini Naidu a/p Ragavelu	1141127226	Kirba4796@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 weather data, time data and the number of bikes recorded to be on the move, categorized in casual or registered users.

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.

The attributes in the dataset are described as follows:

1) **instant:** this is the row index of the dataset, representing either one hour recorded, or one day recorded in the case of day.csv
2) **dteday:** This is the date of which the data was recorded, formatted in “YYYY-MM-DD”
3) **season:** This is the season of the time when the data was recorded. It is stored as an integer representing 1 for spring, 2 for summer, 3 for autumn, and 4 for winter.
4) **yr:** The year of the data recorded. As the dataset is only for the year 2011 and 2012, this is stored as 0 for year 2011, and 1 for year 2012.
5) **mnth:** The month of the data recorded, stored as an integer where 1 is January and so on until it is 12 for December
6) **hr:** The hour of the data, from 0 to 23. This attribute is not present in day.csv.
7) **holiday**: This column records whether the day was a holiday for the state of Washington, where the data was recorded.
8) **weekday**: This represents the day of the week for the data, from 0 as Sunday to 6 as Saturday.
9) **workingday**: This shows whether the day was a working day or not.
10) **weathersit**: This shows the description of the weather at the time of the record. This is stored as an integer, categorizing different weather description into four categories:
    - 1 = Clear, Few clouds, Partly cloudy, Partly cloudy
    - 2 = Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
    - 3 = Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
    - Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
11) **temp:** This is the temperature recorded at the time. In both datasets, this value is normalized according to a certain equation.

12) **atemp:** This is the feeling temperature recorded at the time. Like temp, this value is normalized
13) **hum:** This is the recorded humidity of the data. This value is normalized.
14) **windspeed:** This is the recorded windspeed of the data. This value is normalized.
15) **casual:** This records the count of casual users that have rented a bike and is on the move.
16) **registered:** This records the count of registered users that have rented a bike and is on the move
17) **cnt:** This is the sum of casual and registered users.

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 characteristic leads to various levels of mobility, represented by 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.

DATA PREPROCESSING

Refer to preprocess.R inside the BIKESHARE folder for all of the data preprocessing steps

Various preprocessing steps can be performed on the dataset, whether to verify certain attributes, or to extract certain information.

CHECK OF CNT

According to the original source, the variable **cnt** is the sum of casual and registered users. With R, we can check whether that is true with the following code fragment:

```
# cnt check
# ensure that cnt is the sum of casual and registered
indicesD <- which(datasetD$casual + datasetD$registered != datasetD$cnt)
indicesH <- which(datasetH$casual + datasetH$registered != datasetH$cnt)
```

By doing this, we found that there are no rows where this rule is violated. Therefore, we can safely ignore casual and registered attribute for the purposes of data mining for the counting of bikes in a row.

CHECK MEAN BASED ON SEASON AND YEAR

Now, we can check the mean of bicycle count, separated by year and season with the following code fragment:

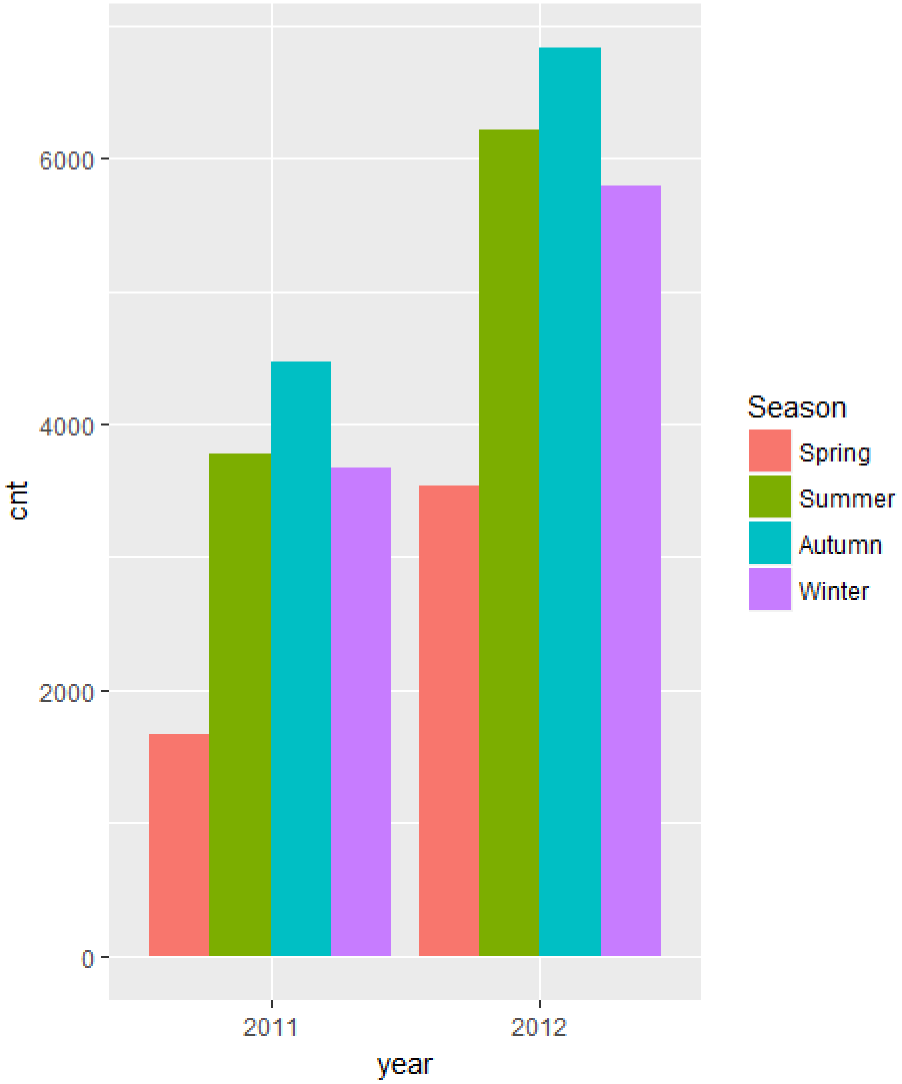
```
library(plyr)
library(ggplot2)

# cnt check
# ensure that cnt is the sum of casual and registered
indicesD <- which(datasetD$casual + datasetD$registered != datasetD$cnt)
indicesH <- which(datasetH$casual + datasetH$registered != datasetH$cnt)

# mean of bike count, sorted by season and time of day
# using the plyr package
bikeMean <- ddply(datasetD, .(datasetD$yr, datasetD$season), summarize,
cnt=mean(cnt))
colnames(bikeMean)[1:2] = c('year','season') # naming columns
bikeMean <- bikeMean[ order(bikeMean$year, bikeMean$season),]
bikeMean$year = as.factor(bikeMean$year)
bikeMean$season = as.factor(bikeMean$season)

# display bar graph
ggplot(bikeMean, aes(year,cnt)) +
  geom_bar(aes(fill = as.factor(season)), position = "dodge", stat="identity")
+
  scale_fill_discrete(name = "Season",
                      labels = c('Spring','Summer','Autumn','Winter')) +
  scale_x_discrete(labels = c('2011','2012'))
```

The image below shows the resulting bar graph:



Here, we can see that despite the cold, winter seasons see more bicycles compared to the spring season. This could be because during the winter, cars would be slowed down, and roads could freeze over and require snow-ploughing before it can be used. Therefore, more commute is done with bicycles.

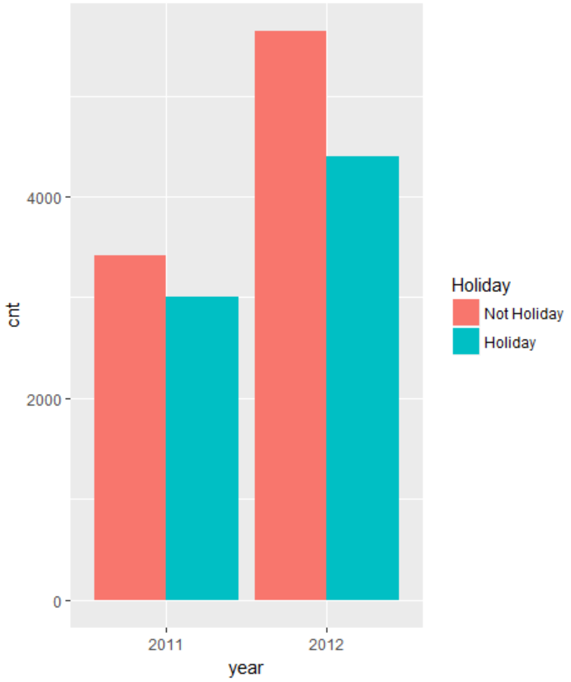
HOLIDAYS AND BIKE COUNT

The next time factor we can consider is whether holidays affect the number of bikes in transit. Using the code below:

```
# mean of bike count, seprate by year and holiday
bikeMean1  <-  ddply(datasetD,  .(datasetD$yr,datasetD$holiday),  summarize,
cnt=mean(cnt))
colnames(bikeMean1)[1:2] = c('year','holiday')
bikeMean1 <- bikeMean1[ order(bikeMean1$year, bikeMean1$holiday),]
bikeMean1$year = as.factor(bikeMean1$year)
bikeMean1$holiday = as.factor(bikeMean1$holiday)

# bar graph plot
ggplot(bikeMean1, aes(year,cnt)) +
      geom_bar(aes(fill  =  as.factor(holiday)),  position  =  "dodge",
stat="identity") +
  scale_fill_discrete(name = "Holiday",
                      labels = c('Not Holiday','Holiday')) +
  scale_x_discrete(labels = c('2011','2012'))
```

We get the following image:



Holidays see lower count of bicycles. One easy inference is that because it is a holiday, people may want to leave the city on a trip, lowering the number of citizens in a city resulting in lower mobility.

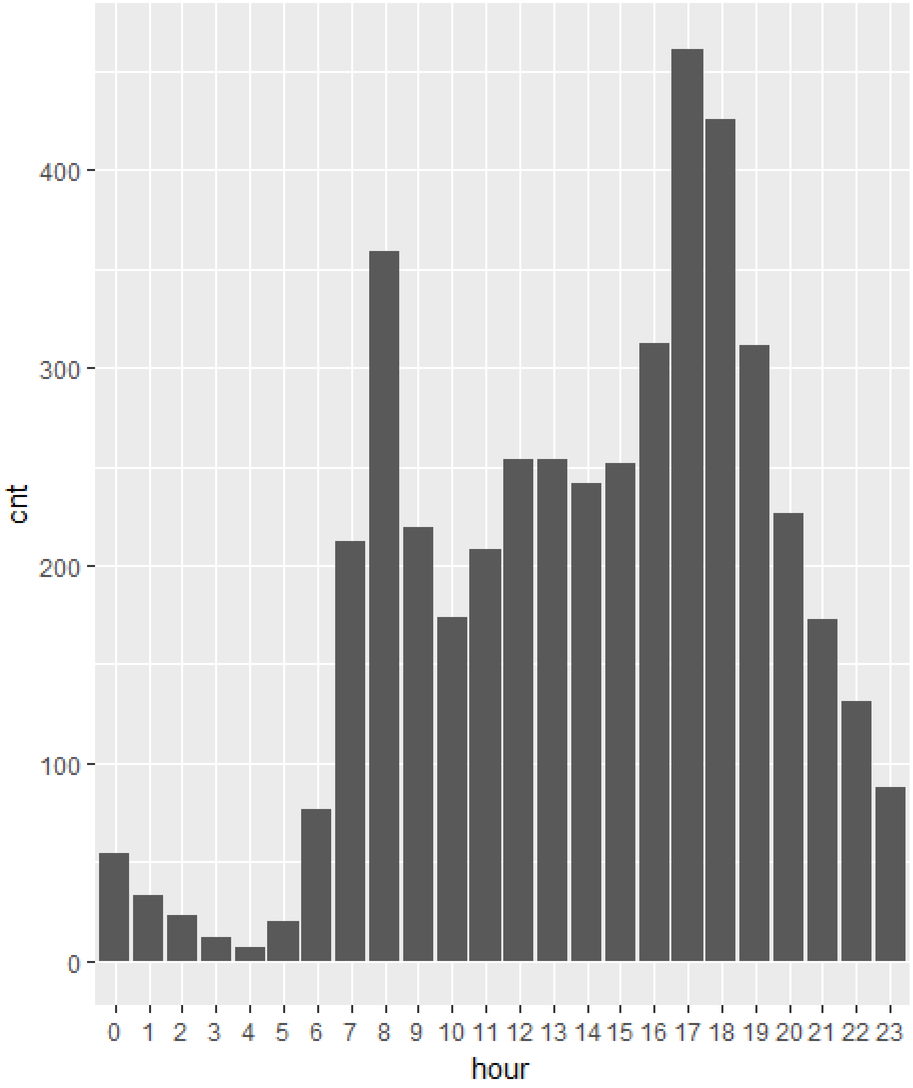
IDENTIFYING PEAK HOURS

Using the hour.csv dataset, we can obtain the mean of bike count grouped by the hour in which it was recorded. By using the follow R code fragment:

```
# mean of bike count, separate by hours in a day
# must use hour.csv for this
bikeMean3 <- ddply(datasetH, .(datasetH$hr), summarize, cnt=mean(cnt))
colnames(bikeMean3) <- c('hour','cnt')
bikeMean3$hour <- as.factor(bikeMean3$hour)

# bar graph
ggplot(bikeMean3, aes(hour,cnt)) +
  geom_bar(stat="identity")
```

We get the following graph and can identify the peak hours:



DATASET VISUALS

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

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

The visuals from this report can be viewed from this app, and you can easily switch between relevant charts. There is also a customizable scatter plot where you can select data based on year and season, and pick attributes as x-axis or y-axis.