# Stat 440 Group Project Report
# Spring, 2015

# Data Management for Bike Sharing Markets in Washington D.C.

Group 17

Haiqian Wang
Yang Xu
Yufei Gui
Yuting Zheng

# Contents

# Introduction

The project that our team works on is the data management on bike sharing markets in Washington D.C. from 2011 to 2012. The data set that we used is the data set from UCI machine learning repository. Two data sets, Day.csv and Hour.csv, are collected from Capital Bankshare System, Washington D.C., USA. They record from 2011 to 2012 and have 16 variables in common with exception that variable "hr" only shows in Hour.csv. Most variables are related to weather, date, season and etc. since they are influential to the renting behavior. Hour.csv contains 17379 observations aggregated by two-hour basis while Day.csv contains 731 observations aggregated by daily basis.

The research interest in this data set primarily focus on the significant meaning of the bike rental system to the real world transportation system. Nowadays, more a more people decide to choose bikes instead of private cars or public transportation. For research interest in economy aspects, as the rental bike system becoming more prosperous, they provide more jobs for the local people. For the possible research interest in environment aspect, since bike is a kind of green transportation, people could research the relationship between the increases in the use of bikes with different environment index. For the transportation aspects, the research interest could concentrate on whether the rental bike system could have positive impacts on relieving the heavy transportation pressure.

The purpose of our project is to prepare and analyze the data to explore the potential factors that will influence the bike sharing markets. Moreover, we also research the possible trend of bike sharing markets from 2011 to 2012.

Note: Column B in both files must have their width enlarged to read in SAS correctly, that is eliminating "#" symbol.

## *Attributes of day data set*

| | | | |
|---|---|---|---|
| **Data Set Name** | PROJECT.DAY | **Observations** | 731 |
| **Member Type** | DATA | **Variables** | 16 |
| **Engine** | V9 | **Indexes** | 0 |
| **Created** | 05/11/2015 20:58:08 | **Observation Length** | 128 |
| **Last Modified** | 05/11/2015 20:58:08 | **Deleted Observations** | 0 |
| **Protection** | | **Compressed** | NO |
| **Data Set Type** | | **Sorted** | NO |
| **Label** | | | |
| **Data Representation** | WINDOWS_64 | | |
| **Encoding** | wlatin1 Western (Windows) | | |

## *Attributes of hour data set*

| | | | |
|---|---|---|---|
| **Data Set Name** | PROJECT.HOUR | **Observations** | 17379 |
| **Member Type** | DATA | **Variables** | 17 |
| **Engine** | V9 | **Indexes** | 0 |
| **Created** | 05/11/2015 20:58:08 | **Observation Length** | 136 |
| **Last Modified** | 05/11/2015 20:58:08 | **Deleted Observations** | 0 |
| **Protection** | | **Compressed** | NO |
| **Data Set Type** | | **Sorted** | NO |
| **Label** | | | |
| **Data Representation** | WINDOWS_64 | | |
| **Encoding** | wlatin1 Western (Windows) | | |

# Methods

The methods that we applied are listed below:

1. Read the raw data files: day.csv and hour.csv related to our project;
2. Check and clean data including bad observations;
3. Name labels to different variables;
4. Set formats for variables;
5. Derive new variables by calculating related existed variables;
6. Subset data;
7. Merge Data;
8. Use freq and means descriptive statistics like mean, maximum and minimum to help us make a more comprehensive conclusion.

# Data Checking

## Check missing values

***No missing values in day data set***

**The FREQ Procedure**

| Number of Variable Levels | | |
|---|---|---|
| **Variable** | **Label** | **Levels** |
| **instant** | Instant | 731 |
| **date** | Date | 731 |
| **season** | Season | 4 |
| **year** | Year | 2 |
| **month** | Month | 12 |
| **holiday** | Holiday | 2 |
| **weekday** | Weekday | 7 |
| **workingday** | Working day | 2 |
| **weathersit** | Weather Situation | 3 |
| **temp** | Normalized Temperature in Celsius | 499 |
| **atemp** | Normalized feeling temperature in Celsius | 690 |
| **hum** | Normalized humidity | 595 |
| **windspeed** | Normalized wind speed | 650 |
| **casual** | Count of casual users | 606 |
| **registered** | Count of registered users | 679 |
| **cnt** | Count of total rental bikes | 696 |

# Check missing values

***No missing values in hour data set***

***The FREQ Procedure***

| Number of Variable Levels | | |
|---|---|---|
| **Variable** | **Label** | **Levels** |
| **instant** | Instant | 17379 |
| **date** | Date | 731 |
| **season** | Season | 4 |
| **year** | Year | 2 |
| **month** | Month | 12 |
| **hour** | Hour | 24 |
| **holiday** | Holiday | 2 |
| **weekday** | Weekday | 7 |
| **workingday** | Working day | 2 |
| **weathersit** | Weather Situation | 4 |
| **temp** | Normalized Temperature in Celsius | 50 |
| **atemp** | Normalized feeling temperature in Celsius | 65 |
| **hum** | Normalized humidity | 89 |
| **windspeed** | Normalized wind speed | 30 |
| **casual** | Count of casual users | 322 |
| **registered** | Count of registered users | 776 |
| **cnt** | Count of total rental bikes | 869 |

We found that there are no missing values in two data sets by using proc freq.

# Data checking by Holiday

| Obs | Date | Holiday |
|---|---|---|
| 17 | 01/17/2011 | holiday |
| 52 | 02/21/2011 | holiday |
| 105 | 04/15/2011 | holiday |
| 150 | 05/30/2011 | holiday |
| 185 | 07/04/2011 | holiday |
| 248 | 09/05/2011 | holiday |
| 283 | 10/10/2011 | holiday |
| 315 | 11/11/2011 | holiday |
| 328 | 11/24/2011 | holiday |
| 360 | 12/26/2011 | holiday |
| 367 | 01/02/2012 | holiday |
| 381 | 01/16/2012 | holiday |
| 416 | 02/20/2012 | holiday |
| 472 | 04/16/2012 | holiday |
| 514 | 05/28/2012 | holiday |
| 551 | 07/04/2012 | holiday |
| 612 | 09/03/2012 | holiday |
| 647 | 10/08/2012 | holiday |
| 682 | 11/12/2012 | holiday |
| 692 | 11/22/2012 | holiday |
| 725 | 12/25/2012 | holiday |

2011 Holiday Schedule

Monday, January 17, 2011 Martin Luther King Jr. Day
Monday, February 21, 2011 Washington's Birthday
Friday, April 15, 2011 DC Emancipation Day*
Monday, May 30, 2011 Memorial Day
Monday, July 4, 2011 Independence Day
Monday, September 5, 2011 Labor Day
Monday, October 10, 2011 Columbus Day
Friday, November 11, 2011 Veterans Day
Thursday, November 24, 2011 Thanksgiving Day
Monday, December 26, 2011 Christmas Day**

2012 Holiday Schedule

Monday, January 2, 2012 New Year's Day*
Monday, January 16, 2012 Martin Luther King Jr. Day
Monday, February 20, 2012 Washington's Birthday
Monday, April 16, 2012 DC Emancipation Day
Monday, May 28, 2012 Memorial Day
Wednesday, July 4, 2012 Independence Day
Monday, September 3, 2012 Labor Day
Monday, October 8, 2012 Columbus Day
Monday, November 12, 2012 Veterans Day*
Thursday, November 22, 2012 Thanksgiving Day
Tuesday, December 25, 2012 Christmas Day

In this section, we check whether the holiday listed in our SAS data set matches the 2011 and 2012 holiday schedule. In this procedure, we create a PROC print procedure and use where to subset the observations that are marked as holiday. The result of this check is that all the holidays in the SAS data set matches the actual holiday schedules. So there is no problem on holiday variable.

# Check total rental counts differences from two data sets

```
2399  proc print data=compare_sum;
2400    title 'Check difference in total counts between hour.csv and day.csv';
2401    where sum_cnt~=cnt;
2402  run;

NOTE: No observations were selected from data set WORK.COMPARE_SUM.
NOTE: There were 0 observations read from the data set WORK.COMPARE_SUM.
      WHERE sum_cnt not = cnt;
NOTE: PROCEDURE PRINT used (Total process time):
      real time           0.00 seconds
      cpu time            0.00 seconds
```

We use hour data set to calculate the accumulated values of total counts of rental bikes for each day. By comparing the accumulated values named "sum_cnt" with the total counts named "cnt" in day data set, we have found that there is no difference between these two variables, which indicates that researchers make no mistake in calculating.

# Descriptive statistics of data set

***Descriptive statistics of day data set***

***The MEANS Procedure***

| Variable | Label | N | Mean | Std Dev | Minimum | Maximum |
|---|---|---|---|---|---|---|
| temp | Normalized Temperature in Celsius | 731 | 0.4953848 | 0.1830510 | 0.0591304 | 0.8616670 |
| atemp | Normalized feeling temperature in Celsius | 731 | 0.4743540 | 0.1629612 | 0.0790696 | 0.8408960 |
| hum | Normalized humidity | 731 | 0.6278941 | 0.1424291 | 0 | 0.9725000 |
| windspeed | Normalized wind speed | 731 | 0.1904862 | 0.0774979 | 0.0223917 | 0.5074630 |
| casual | Count of casual users | 731 | 848.1764706 | 686.6224883 | 2.0000000 | 3410.00 |
| registered | Count of registered users | 731 | 3656.17 | 1560.26 | 20.0000000 | 6946.00 |
| cnt | Count of total rental bikes | 731 | 4504.35 | 1937.21 | 22.0000000 | 8714.00 |

***Descriptive statistics of hour data set***

***The MEANS Procedure***

| Variable | Label | N | Mean | Std Dev | Minimum | Maximum |
|---|---|---|---|---|---|---|
| temp | Normalized Temperature in Celsius | 17379 | 0.4969872 | 0.1925561 | 0.0200000 | 1.0000000 |
| atemp | Normalized feeling temperature in Celsius | 17379 | 0.4757751 | 0.1718502 | 0 | 1.0000000 |
| hum | Normalized humidity | 17379 | 0.6272288 | 0.1929298 | 0 | 1.0000000 |
| windspeed | Normalized wind speed | 17379 | 0.1900976 | 0.1223402 | 0 | 0.8507000 |
| casual | Count of casual users | 17379 | 35.6762184 | 49.3050304 | 0 | 367.0000000 |
| registered | Count of registered users | 17379 | 153.7868692 | 151.3572859 | 0 | 886.0000000 |
| cnt | Count of total rental bikes | 17379 | 189.4630876 | 181.3875991 | 1.0000000 | 977.0000000 |

***No specific demand for bike sharing***

| Obs | date | cnt |
|---|---|---|
| **1** | 09/15/2012 | 8714 |
| **2** | 09/29/2012 | 8555 |
| **3** | 09/22/2012 | 8395 |
| **4** | 03/23/2012 | 8362 |
| **5** | 05/19/2012 | 8294 |
| **6** | 09/09/2012 | 8227 |
| **7** | 07/25/2012 | 8173 |
| **8** | 09/21/2012 | 8167 |
| **9** | 10/05/2012 | 8156 |
| **10** | 06/02/2012 | 8120 |

According to the statistics in the means table, we found the largest value of total count of rental bikes in the “day.csv” is 8714 but the smallest value is 22 per day. So we double the whether the largest value is reliable. Then we sort the “day.csv” by descending order and then print the largest 10 values of the data set. Finally, we found that the largest 10 values are all in the range of 8000-9000 so we conclude that there is no problem on the value of total counts of rental bikes.

# Results

## Compare total counts by season for 2 years

| Obs | Season | Total counts of rental bikes in 2011 | Total counts of rental bikes in 2012 | Increase | Increase rate |
|---|---|---|---|---|---|
| **1** | Spring | 150000 | 321348 | 171348 | 114% |
| **2** | Summer | 347316 | 571273 | 223957 | 64% |
| **3** | Fall | 419650 | 641479 | 221829 | 53% |
| **4** | Winter | 326137 | 515476 | 189339 | 58% |

**Analyzed variables: season, total rental counts in 2011 and 2012.**

In order to compare the counts of total rental bikes among different seasons for two years, we sort data by year and season, and then create two data sets by calculating the accumulated values for each season in 2011 and 2012. After that, we merge these two data sets and create two new variables named ‘increase’ and ‘increase rate’, which make it clear to see changes in total counts of rental bikes for each season between two years.

Therefore, it is obvious that there is a significant increase in bike sharing from 2011 to 2012 for the same season, especially for Spring. According to the result the table shows, we can draw the conclusion that people have a greater demand in bike rental in 2012, which indicates that an increasing number of people would like to accept the new rental market, regardless of seasons. Specifically, the number of bike-users reach the maximum in Fall for its comfortable weather in Washington D.C. in Fall.

# Compare total counts for casual users and registered users

| Year | Total rentals of casual users | Total rentals of registered users | Total rental | Percentage of rental of casual users | Percentage of rental of registered users |
|---|---|---|---|---|---|
| 2011 | 247252 | 995851 | 1243103 | 20% | 80% |
| 2012 | 372765 | 1676811 | 2049576 | 18% | 82% |

**Analyzed variables: rentals of casual users, rentals of registered users and total rentals.**

We found that users are categorized as casual users and registered users thus it is worth looking into it. We sort data by year and create three variables tot_casual (total rental counts of casual users), tot_registered (total rental counts of registered users) and total (rental counts in total). At the last observation of each year, we calculate the percentage of rental of casual users and percentage of rental of registered users.

Observing the result from table above, majority of total rentals are from registered users. It implies that rental services have quite loyal registered users and they probably are citizens of Washington D.C. Tourists may contribute to minority of rental counts from casual users. Also total rental counts increased by 65% which indicates that people are aware of benefits of riding bikes.

# Compare rental counts on working and non-working days

| Total counts on working days | Total counts on holiday or weekend | Counts in total | Percentage of rental counts on holiday or weekend | Percentage of rental counts on working days |
|---|---|---|---|---|
| 2292410 | 1000269 | 3292679 | 30% | 70% |

**Analyzed variables: rental counts on working days and nonworking days.**

In this section, we create three variables, work_count (total rentals on working days), nonwork_count (total rentals on non-working days) and total (rentals in total). We found that about 70% of rentals happen on working days which implies that citizens of Washington D.C may ride bike to work or school. Therefore rental service plays a quite important role to their daily life.

# Analysis of total count of rental bikes under different weather situation from 2011 to 2012

| Obs | Weather Situation | Total count of rental bikes under different weather situation in 2011 | Total count of rental bikes under different weather situation in 2012 | diff | Increase rate |
|---|---|---|---|---|---|
| **1** | Clear | 835067 | 1422885 | 587818 | 70% |
| **2** | Mist & Cloudy | 382924 | 613934 | 231010 | 60% |
| **3** | Light Precipitation | 25112 | 12757 | -12355 | ( 49%) |

**Analyzed Variables: Weather Situation, Total Counts of Rental Bikes in different years**

On the programming part, first, we read in the raw data into SAS. We sort the data by year and weather situation. And then we create two data set to display the total count of rental bikes in 2011 and 2012 by the method of conditional output. In this data step, we used by-statement and first and last variables to calculate the total rental bikes in 2011 and 2012. After creating and output the two data sets of total rental bikes, we applied a merge of these two sets by the weather situation variable. Next, we create a new variable called difference to hold the difference between the total count of rental bikes in 2011 and 2012 under three different weather condition. Moreover, we also create a new variable called increase rate to indicate the trend of the total rental bikes amount.

As for the results in the table, under the clear weather situation, there is an increase between 2011 and 2012. The total count of rent bikes in 2012 are 70%more than 2011. Similarly, the total count in 2012 under Mist & Cloudy weather is 60% more than 2011. However, under light precipitation weather, there is a dramatic decrease in 2012, which is 49% less than 2011. According to the results above, we can conclude that more people choose to rent bikes in 2012 than 2011 in the days that is not raining. This indicates that more and more people realize that bike is a more environmental-friendly transportation and riding bikes is also good for health. Moreover, it also indicates that the bike sharing market becomes more well-organized and well-developed.

# Analysis of total count of rental bikes under different time of the day from 2011 to 2012

| Obs | Time period | Total counts of rental bikes in 2011 | Total counts of rental bikes in 2012 | Increase | Increase rate |
|---|---|---|---|---|---|
| **1** | Morning | 119905 | 195907 | 76002 | 63% |
| **2** | Night | 590527 | 960190 | 369663 | 63% |
| **3** | Noon | 532671 | 893479 | 360808 | 68% |

**Analyzed Variables: time, total counts in different years.**

We divided the data into three parts as followings: we name a new variable as 'time' with three separate values as 'Morning', 'Noon' and 'Night', regarding to different ranges of hour as [0, 8), [8, 16) and [16, 23] respectively. We sort the data in hour.csv by year and time, knowing they are highly correlated, but not perfectly correlated. Based on our result, we have our total counts of rental bikes in 2012 more than the total counts of rental bikes in 2011 for all three time periods. The range for increase rate from 2011 to 2012 is 63%-68%, which is quite high, and the highest increase rate occurs in Noon. We think couple reasons for the distribution of increase rate is the advertisement for the bike rental shows a better feedback and more people try to rent bikes for transaction. Moreover, we also found that even the most increase rate turns out in Noon, the most increase amount for bikes rental occurs in Night, this show that more people use bike after their daily work in Night. To sum up, the amount for bikes rental increased a lot during these two years.

# Conclusion

According to the results we described above, we have several conclusions.

- More people tends to rent bikes in fall in general.
- More people choose to become a registered users, which in some way indicates that the bike sharing market is in a good trend of developing.
- More people rent bikes on working days comparing with non-working days.
- Most people prefer renting bikes on clear days.
- Many people choose to rent bikes during the periods of noon and night.
- More people rent bikes in 2012 than 2011.

# Discussion

Evaluating the whole project, there are several points that we believe are the limitations of the analysis.

- In the "hour.csv", there are lots of days that do not have the observations of all 24 hours.
- The data only contains the observations in 2011 to 2012, which makes the results and the conclusions not that convincing.
- There might be some factors that may not be reflected on the data we collected, such as the government policy, which may influence the results of the analysis.

# Reference

Fanaee-T, Hadi, and Gama, Joao, "Event labeling combining ensemble detectors and background knowledge", Progress in Artificial Intelligence (2013): pp. 1-15, Springer Berlin Heidelberg, doi:10.1007/s13748-013-0040-3.

Data source: https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset

Original Source: http://capitalbikeshare.com/system-data

Weather Information: http://www.freemeteo.com

Holiday Schedule: http://dchr.dc.gov/page/holiday-schedule