INFO 5100
PROJECT 3 FINAL REPORT

Bike Sharing Demand in DC

Project Team
- Jiabin Dong(jd836)
- Jing Zhang(jz322)
- Wenhan Xue(wx62)

TA
- Neil Parker(nwp7)

Data Sets

Sources:

The dataset comes from Kaggle Bike Sharing Demand competition dataset. It covers two years bike usage from January 1, 2011 to December 31, 2012. There is a train dataset contains information in first 19 days in each of the month, and a unlabeled test data contains information in the remaining days in that month. In each day, there are 24 items, each representation the usage in one hour. In this project, we use the training dataset, which provides 10,886 observations.
Link: https://www.kaggle.com/c/bike-sharing-demand/data

Variables:

- Datetime: the date, and specific hour (time)
- Season: spring, summer, fall, winter (categorical)
- Holiday: whether a day is a holiday or not (binary)
- Workingday: whether a day is neither a weekend nor holiday (binary)
- Weather: four situations, from clear to extremely bad weather (categorical)
- Temp: in Celsius (continuous)
- Atemp: feels like temperature in Celsius (continuous)
- Humidity: (continuous)
- Windspeed: (continuous)
- Casual: users can be registered to non-registered users. This part is the number of non-registered rentals (continuous)
- Registered: number of registered user rentals (continuous)
- Count: number of total rentals (continuous)

Criteria for Data & Variable Selection

- Model Building: We aimed at choosing a dataset could combine both description of data and some sort of statistical models.
- Influence of the Usage: We chose variables could have significant influence on the bike demand. This is done by the assumption in real life and preliminary statistical findings in R by Random Forest Tree.

- Variety of Data Type: We hope to contain both categorical and continuous independent variables in our website.

Description of Mapping to Visual

Overall Design:

- Make the instruction friendly for users
- Choose appropriate interaction method (e.g Drop-down menu, slider bar)
- Keep the text consistent with each other
- Highlight key elements in different font size and color
- Data transformation: the time is parsing according to format we need, and some variables are changed to different names in order to show clearly in the website

Part1:

Weekly Animation(By Average)

- Left graph: Shows average count number in the selected period. Each bicycle image stands for ten bicycle rentals. Pack layout is used here for display the multiple bicycles.
- Right graph: The trend through a day in different weekdays and weekends. This graph helps users get a general idea of the how bike demand change over a day. Both x-axis and y-axis are linear scaled.
- Input box: User can choose start time and end time of bike usage
- Slider bar: Specify the time in the period, to see the average count numbers
- Control panel: Choose each day in a week to see the trend. The control panel connects the left graph and right graph. We're expecting different trend between weekdays and weekend.

Part2:

Bike Rentals Trends in 2011~2012

- Each scatter is a circle showing the numbers of bicycle usage
- X-axis is the month, and each bar represents usage in the month
- Y-axis is the number of bike demand
- Both axises apply the linear scale
- The color of the circle represents the temperature. As the temperature goes up, the color turns into red and into yellow as the opposite. The legend is given at the left top corner of the graph

Part3:

What Could Impact Bike Rentals?

- Average graph: Each point represents the mean in a specific value of the selected factor
- Scatter graph: Each point is an observation
- X-axis is the factor(temperature, humidity, wind speed)
- Y-axis is the number of bike demand
- Both axises apply the linear scale
- Model: Linear model using least squares

- Brushing: Specify the data points, and the model is rebuilt using the data in the range

Part4:
Now, Season and Weather are considered together

- Average graph: Each point represents the mean in that season and weather situation
- Scatter graph: Each point is an observation at a random position bounded by the area of corresponding season and weather
- Random Scatter graph: Each point is an observation in a random space in the graph
- X-axis is the season
- Y-axis is the weather situation
- Both the axises use the ordinal scale. Color scale represents the count of bike rentals. More darker green, more bike demand.
- Color legend is given at the right top corner

Plotted but not used:
Bike sharing in four seasons
In this graph, we plotted the trend in a same season, and separate the numbers in stacks with registered and casual users. Readers could choose different season to see the trend. The graph is not used at last, because: 1) There is some overlapping information between this graph and the overall trend graph. 2) There are gaps between some days for the reason of testing data, and aesthetically it is not satisfactory.
Play with log numbers and other variables such as humidity
We also tried to plot the temperature vs count models by log(numbers), but the trend is not clear.

The Story

- Finding1: The number changes with the hours in a day, and it is conditional to working day. There is one peak in weekends, and the peak is at the noon. There are two peaks(at the commuter time) in weekdays, and approximated the rush hours in the morning and night.
- Finding2: In the two years data, we could see an increasing trend across the time span. Probably bicycle rent is getting popular!
- Finding3: The temperature is correlated with the month in the year, which is reasonable. There is a linear relationship between temperature and average bike demand. In addition, humidity also impacts the number of bike rentals with negative correlation. We cannot tell too much whether the wind speed is also a reason from the model graph. But obviously, people do not prefer to rent bike in windy days. Further statistical inference is needed to see whether there are significant results.
- Finding4: The usage is near zero in severe weather. The demand is more in summer and fall, and in clear weather.