

### Explanation for QM Project

Task: Predict the amount of bikes rented each hour for the days 16-19. The data are in the test sample ?

To build the model use the train data (days 1-15).

Step by Step Procedure: (make sure you write what you are doing in the report + findings)

Find out which variables are relevant. For each variable do the following:

Do a regression (Table 1)

- 1) Go into train data
- 2) Select data, data analysis, regression
- 3) Select as y the count (M1:M8601 in my case)
- 4) Select as x input the variable of interest (e.g. I1:I8601 for humidity)
- 5) Check box: label if your selected column contains a title (e.g. count)
- 6) Check box: Line Fit plots
- 7) Make sure you select new tabellenblatt)

Analyse the result for this variable

- 8) Look at the  $R^2$  (cell B5): no linear relationship if value is about 0.2 or lower
- 9) Look at the line fit plot whether a linear relationship seems reasonable.

Pivot Table: You might also want to look at the variable via a pivot table:

- 10) Select all the train data in the train sample
- 11) Insert, pivot table (make sure you select "New worksheet")
- 12) Drag the variable you want to know about in the "Rows" section
- 13) Drag count in the "values" section
- 14) Select the arrow next to "values"
- 15) Value Field section, average or sum
- 16) Can use Pivot graph to display if you prefer a graphic representation

Decide which variables are relevant based on the above analysis

Time Analysis

- 17) Do the regression for time

- 18) You will find that Time does not obey linear dependence
- 19) easiest solution: Do separate regressions for each of the 24 time variables.
- 20) If this is too much work: combine times into time buckets according to similar counts.
- 21) Suppose you want the following times into 1 basket: times 0:00:00, 01:00:00, 02:00:00, 04:00:00, 05:00:00, because the count is very similar for all these times and we can do 1 regression for all these times.
- 22) Click any cell in the first row then select filter under data
- 23) In the time column deselect all and check the times you want to include your first time bucket (see 21)
- 24) Copy paste only the selection into a new worksheet: "Ctrl A", "Ctrl C". Add a new worksheet and "Ctrl V".
- 25) Make sure the final variables you want to include are all next to each other
- 26) Rename Worksheet and do a regression in this worksheet as above. Assuming you have 2000 entries the input y range is M1:M2000.
- 27) For the input x range include all the variables you want in your final regression model.
- 28) Example: If atemp and humidity are your independent variables and they are next to each other (have to be) in columns J and K, then input x range is J1:K2000.
- 29) The  $R^2$  value is in the "R square" cell. The intercept and the coefficients are also stated in B17, B18, B19. For the example above: count = intercept + atemp-coefficient\*atemp + humidity-coefficient\*humidity
- 30) Write down the equation and repeat for your other time buckets
- 31) For the prediction (test sample) you have to check (using if statements or a lookup table or split the test sample into the time buckets each a different spreadsheet) which time basket the time belongs to and then use the corresponding formula.
- 32) Some details for the If method:

=IF(OR(HOUR(B2)=0;HOUR(B2)=1;HOUR(B2)=2);"Time bucket 1";"Other time buckets"). This checks which hours are in the first time bucket. Replace "Time bucket 1" with the formula for the first time bucket. Replace other time buckets with further if statements like above

- 33) Some details on Splitting test sample into the corresponding times of your time bucket: insert the formula in 29) for count. Finally you might want to combine the different worksheets into one.

Table 1