Study the distribution of the number of registered users per day (registered). The key features of these distributions.

The MEANS Procedure

Analysis Variable : registered

N	N Miss	Mean	Median	Mode	Std Dev	Minimum	Maximum	Lower Quartile	Upper Quartile	Kurtosis	Skewness	Lower 99% CL for Mean	Upper 99% CL for Mean	Std Error
731	0	3656.172	3662.000	1707.000	1560.256	20.000	6946.000	2493.000	4790.000	-0.713	0.044	3507.136	3805.208	57.708

Figure 1

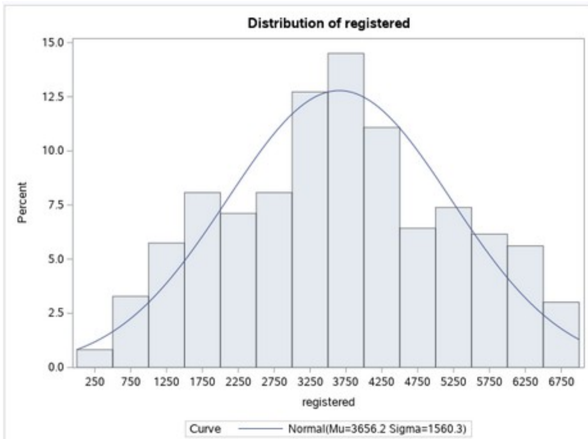


Figure 2

Goodness-of-Fit Tests for Normal Distribution				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	D	0.04316104	Pr > D	<0.010
Cramer-von Mises	W-Sq	0.22789926	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	2.10334142	Pr > A-Sq	<0.005

Figure 3

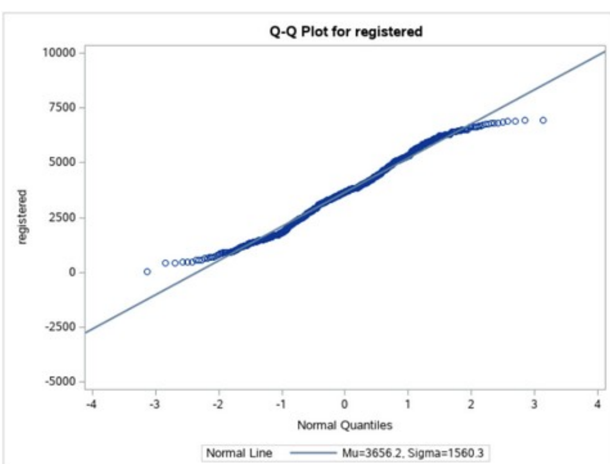


Figure 5

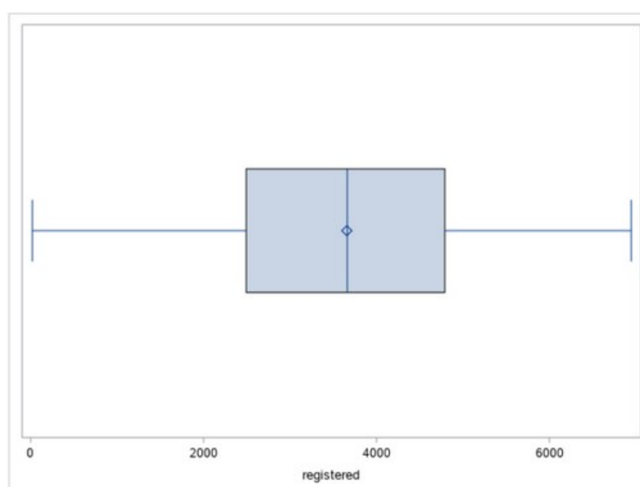


Figure 4

According to figure 1, the mean of data of the registered user is 3656, and we are 95% confident that the population mean of the registered user is between 3507 and 3805 users per day. We have a large sample size (n>30) to move with producing a confidence interval.

The data for the registered user has a negative kurtosis (-0.713) and is slightly right-skewed. The negative kurtosis (light tail) agrees with the histogram. The Q-Q plot in figure 5 implies no deviation from the straight line in the centre, but a negligible deviation on both sides implies a light tail(negative kurtosis). Sometime for skewed distribution we use mathematical transformer. We apply log and sq root transformer to handle skewness, but the result was not satisfying.

As figure 1 shows, the median and mean are near each other, which is evident in the box plot. The length of whiskers in the box plot corresponds to a high dispersion (std=1560).

The Kolmogorov-Smirnov test results in Figure 3 denote that the null hypothesis of Normal distribution is rejected and the data of registered user is assumed to be non_Normal, D(731) = 0.04 pvalue = 0.01< 0.05.

boxplots suggests a lot about the pattern and trend of bike rentals

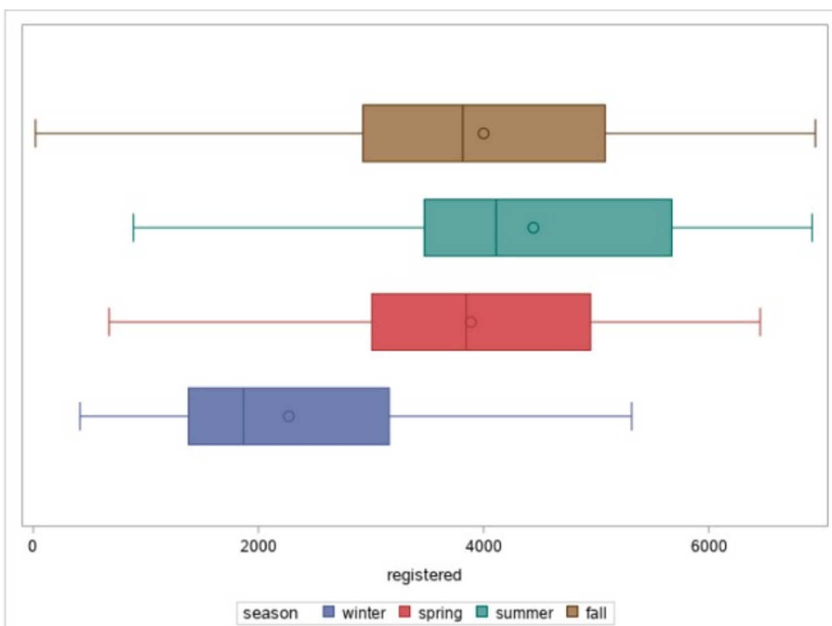


Figure 6

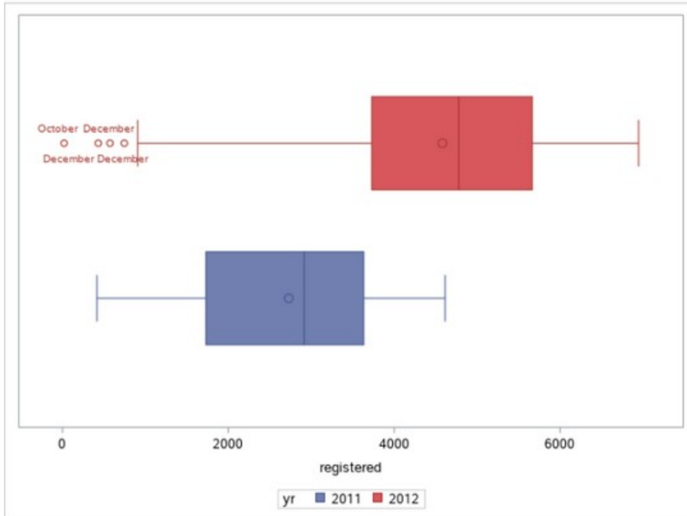


Figure 7

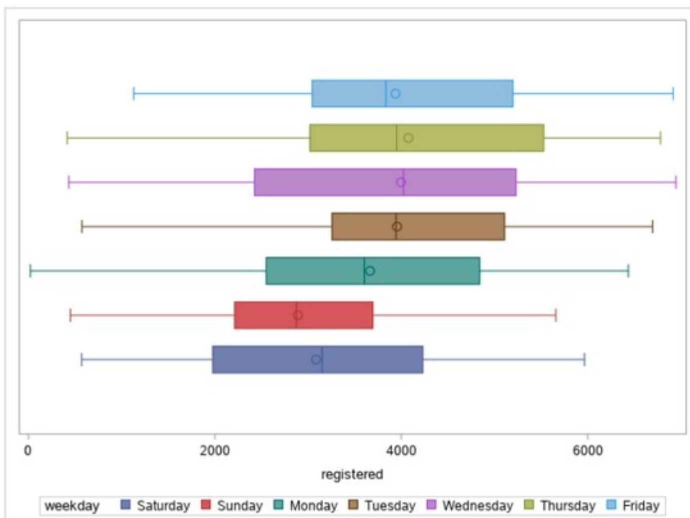


Figure 8

As box plots in figure 7 show, compared with 2011, we can see a surge in registering for the rental bike in 2012. The interquartile(IQR), which implies the dispersion of 50% samples in this year, shifted to the right and accordingly, the mean and median shifted, confirming the increase in registered users in 2012. Furthermore, there are outliers in the box plot of 2012, indicating a sudden decrease in this year. Also, the maximum and minimum count of registered users were recorded in 2012.

Figure 6 shows a pattern of increase in rental bikes. While the records of maximum and minimum registered users for bike rentals happened in fall (cause a long whisker), according to mean and median, more people attend biking in summer, and fewer people attend in winter, demand is at its bottom in winter and increases by spring. This trend continues until it hits the peak in summer as IQR, mean and median shift to the right, and by fall, the fall starts.

Box plots for weekdays indicate a pattern of increase during days. The number of bikes registered on Sunday, as left shifted IQR and the minimum mean and median implies, is at the bottom. The demand for bikes increases, as IQR shifted to the right and mean and median also increases, on Wednesday and Thursday.

In 2012, the east coast of the United States was struck by Hurricane Sandy. There is a severe weather event evident in your results.

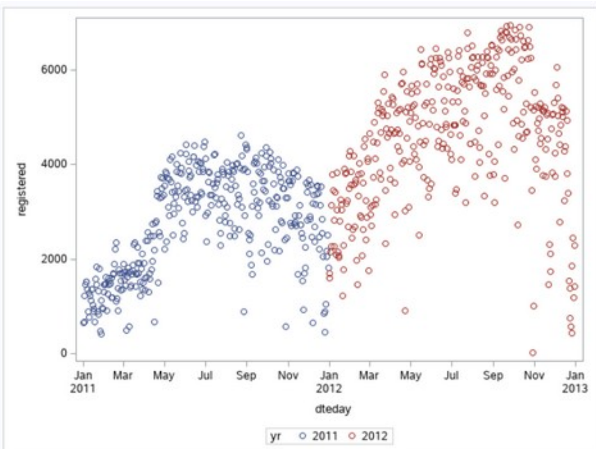


Figure 9

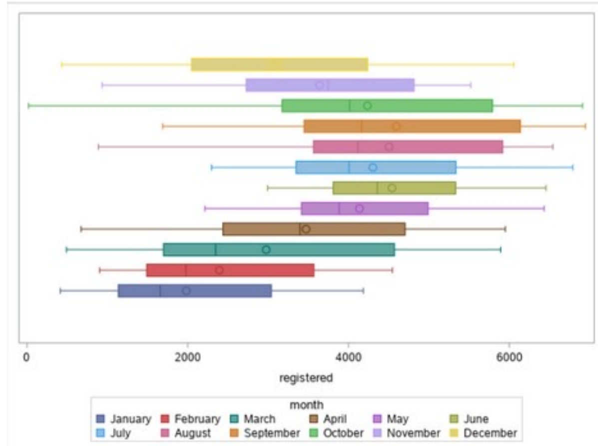


Figure 10

As the box plot of the year 2012 in figure 7 shows, the outlier implies a sudden drop in October and December. Despite the growth of demand in 2012, the minimum demand is recorded this year. Also, by considering box plots for months in figure 10, we notice the minimum demand happened in October. Furthermore, the minimum demand happened on Monday according to weekday box plots. Also, the scatter plot in figure 9 confirms a sudden drop in the incremental trend of registered user in 2012. Overall, as the pattern in box plots and the scatter plot implies, a special weather event is assumed that occurred on Monday, October 2012.

A Pearson correlation matrix relating variables registered, atemp, temp, hum and windspeed. There are relationships that stand out from the rest

Pearson Correlation Statistics (Fisher's z Transformation)									
Variable	With Variable	N	Sample Correlation	Fisher's z	Bias Adjustment	Correlation Estimate	95% Confidence Limits		p Value for H0:Rho=0
registered	atemp	731	0.54419	0.61009	0.0003727	0.54393	0.490773	0.593052	<.0001
registered	temp	731	0.54001	0.60417	0.0003699	0.53975	0.486268	0.589203	<.0001
registered	hum	731	-0.09109	-0.09134	-0.0000624	-0.09103	-0.162468	-0.018636	0.0137
registered	windspeed	731	-0.21745	-0.22098	-0.0001489	-0.21731	-0.285325	-0.147112	<.0001
atemp	temp	731	0.99170	2.74034	0.0006792	0.99169	0.990397	0.992810	<.0001
atemp	hum	731	0.13999	0.14091	0.0000959	0.13989	0.068071	0.210275	0.0001
atemp	windspeed	731	-0.18364	-0.18575	-0.0001258	-0.18352	-0.252673	-0.112505	<.0001
temp	hum	731	0.12696	0.12765	0.0000870	0.12688	0.054869	0.197573	0.0006
temp	windspeed	731	-0.15794	-0.15928	-0.0001082	-0.15784	-0.227746	-0.086313	<.0001
hum	windspeed	731	-0.24849	-0.25380	-0.0001702	-0.24833	-0.315168	-0.179040	<.0001

Figure 11

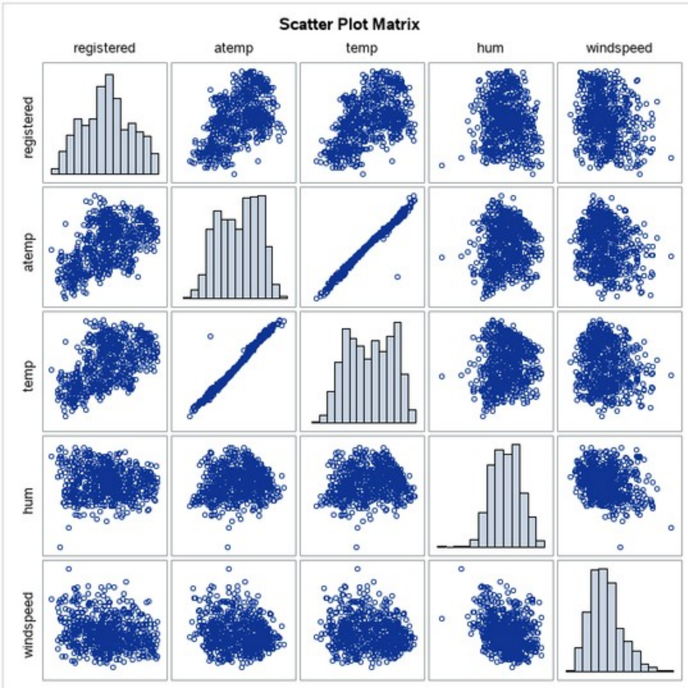


Figure 12

Pearson Correlation Coefficients, N = 731 Prob > |r| under H0: Rho=0					
	registered	**atemp**	**temp**	**hum**	**windspeed**
registered	1.00000	0.54419 <.0001	0.54001 <.0001	-0.09109 0.0138	-0.21745 <.0001
atemp	0.54419 <.0001	1.00000	0.99170 <.0001	0.13999 0.0001	-0.18364 <.0001
temp	0.54001 <.0001	0.99170 <.0001	1.00000	0.12696 0.0006	-0.15794 <.0001
hum	-0.09109 0.0138	0.13999 0.0001	0.12696 0.0006	1.00000	-0.24849 <.0001
windspeed	-0.21745 <.0001	-0.18364 <.0001	-0.15794 <.0001	-0.24849 <.0001	1.00000

Figure 13

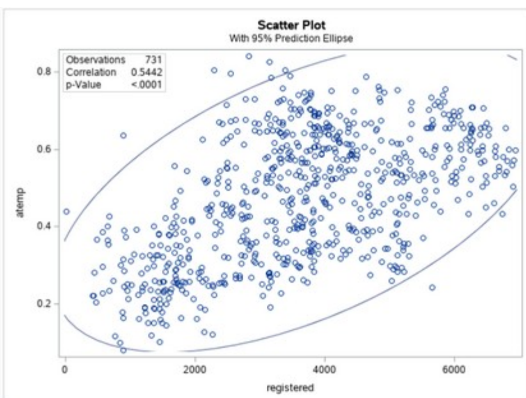


Figure 15

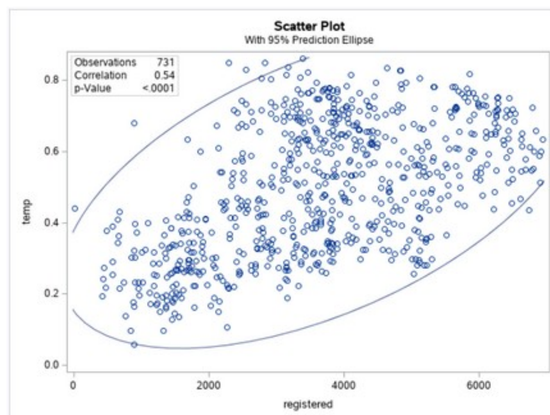


Figure 14

According to figure 13Pearson Correlation Coefficients (PCC) matrix which evaluates the linear relationship between two continuous variables shows that correlation between registered user with atemp and also with temp is strong positive (r = 0.54 also say Larg effect) and is statistically significant (P-value < 0.0001). it means the null hypothesis (r=0 No correlation) is rejected, and the correlation between these variables are significant at 5% level. Accroding to figure 11 we are 95% confident that PCC between these variables in population is between 0.48 and 0.59. As figure 14 and 15 show, the scatter plot between atemp and registered and also temp and registered imply a correlation with somehow large effect, and observations which lying away from the main body (ellipse area) are assumed to be outlier.

PCC between registered user and windspeed implies a medium effect (about - 0.21) and is statistically significant (P-value < 0.0001). We are 95% confident that PCC between these two variables in population is between -0.28 and -0.15. Also, the correlation between registered user and humidity is bout -0.09 which indicate approximately no correlation.

PCC between atemp and temp implies a strong positive correlation with .99 and is statistically significance. We are 95% confident that correlation between atemp and temp in population is between 0.990 and 0.992 and is statistically significant. The scatter plot for these variables in figure 12 shows, there is linear shape implying a strong correlation. As can be seen there is a data error laying away from linear point which affect PCC. The circle shape or curve shape which is apparent in scatter plots in figure 12 between registered with hum and windspeed, and aslo atemp with hum and windspeed, imply a weak correlation. Also, Some outliers are evident in their scatter plots.

A regression model relating registered to atemp, with registered as the dependent variable. Residual plots and influence diagnostics

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	1184.63986	149.20595	7.94	<.0001	891.71525	1477.56448
atemp	1	5210.31247	297.50190	17.51	<.0001	4626.24976	5794.37518

Figure 16

Root MSE	1309.89153	R-Square	0.2961
Dependent Mean	3656.17237	Adj R-Sq	0.2952
Coeff Var	35.82685		

Figure 17

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	526282237	526282237	306.72	<.0001
Error	729	1250829735	1715816		
Corrected Total	730	1777111972			

Figure 18

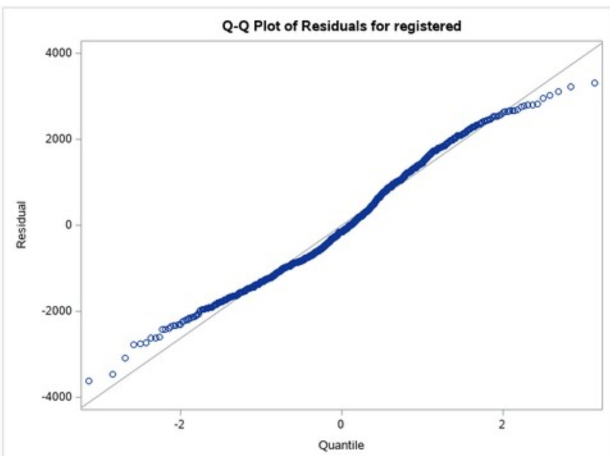


Figure 19

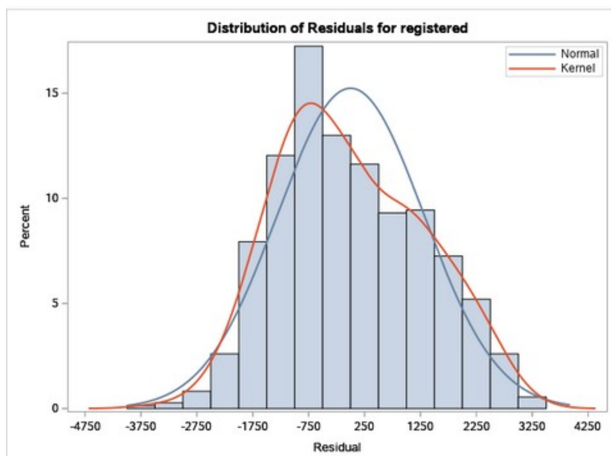


Figure 20

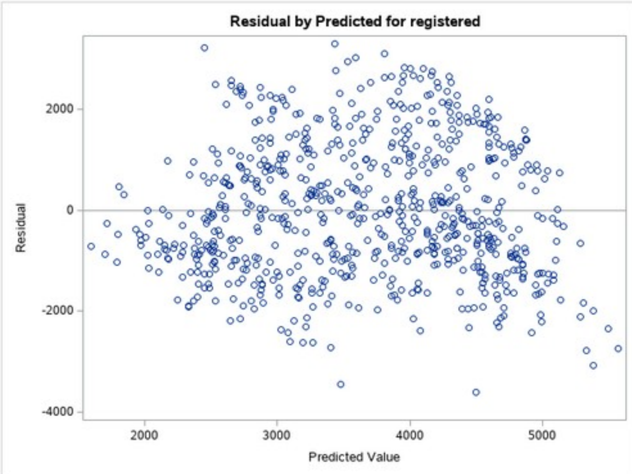


Figure 21

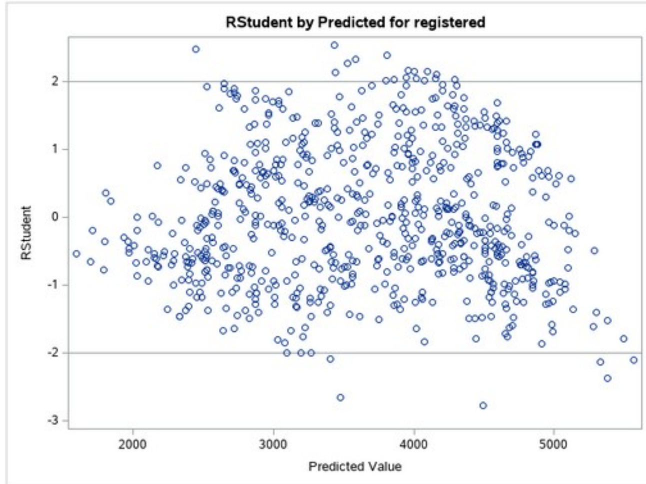


Figure 22

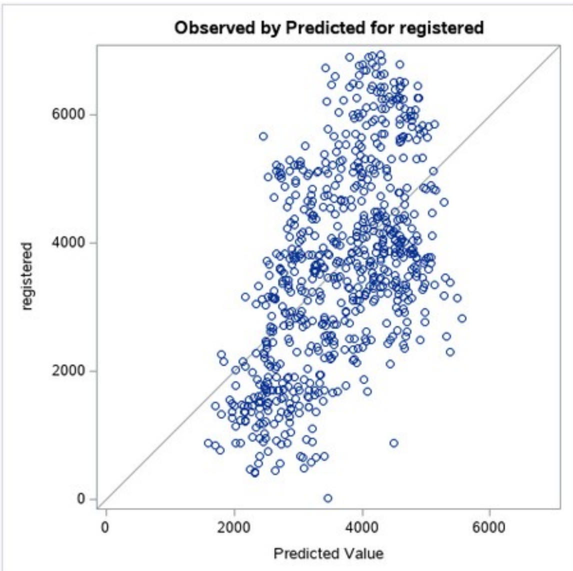


Figure 23

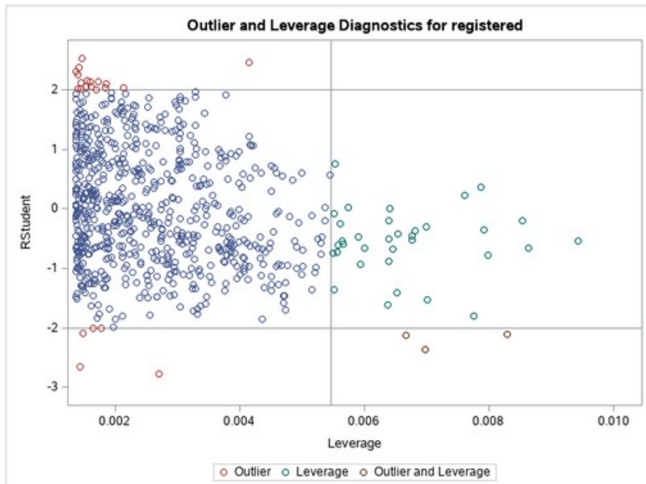


Figure 24

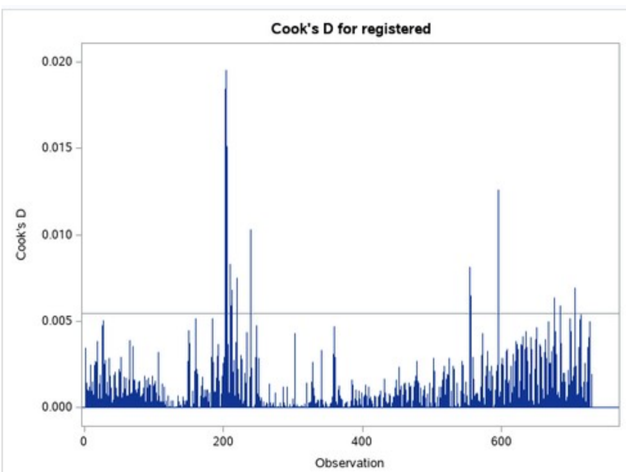


Figure 25

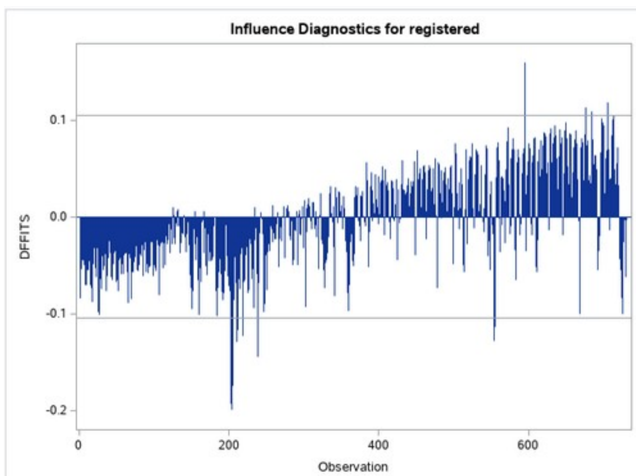


Figure 26

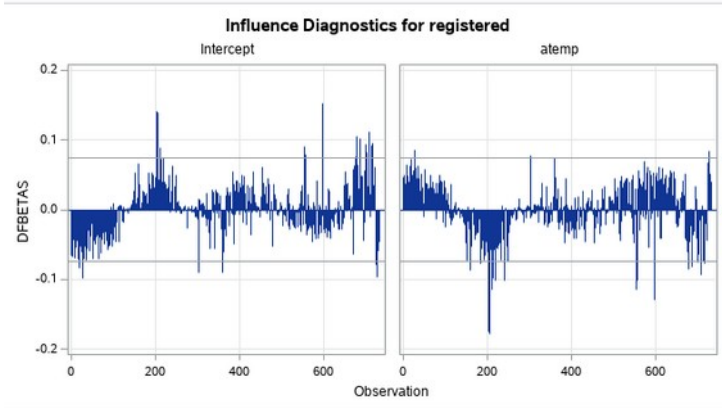


Figure 27

As Analysis of Variance table indicates (figure 18) there is statistically significant relationship between registered and atemp by the large F-ratio (a good model has a large F-ratio). In this model F = 306 and P-value < 0.0001 which indicate the null hypothesis is rejected (H0= no relationship between variables) so the model is statistically significant and we can correctly interpret the result.

As figure 17 shows, the R-square (Coefficient of determination) implies the proportion of variance is about 0.3. It indicates atemp account for about 30% of variance in registered which is a weak strength. It also indicate that atemp explain for about 29% of variance in registered population (Adj R-sq is a bit lower than sample).
The estimated equation is:

$$registered = 1184.64 + 5210.3 \text{ atemp}$$

According to figure 16, the above equation emphasized on the simple linear regression that demonstrates the association of one independent variable (atemp) with continuous dependent variable (registered). As can be seen from the equation, the registered user score with no atemp expected to be 1184.64. Also, the above equation indicated that by 1 unit increase in atemp, registered users increases by 5210.3 implying the positive correlation between registered and atemp.

According to figure 16, the t-statistic for intercept is 7.94 and p-value (<0.0001) confirm the estimation of 1184.64 is statistically significant and also agree with p-value of model. Also, we are 95% confident that the mean of registered users for zero atemp is between 892 and 1478 in population. Also, the t-statistic for slope is 15.51 and p-value <0.0001 ,confirming the slop estimate of 5210.3 is statistically significant. We are also 95% confident that for each 1 unit increase in atemp, the mean of registered users increase between 4626 and 5794 unit in population.

As for the fit diagnostic, four assumptions of linear regression should be taken into consideration. If these four linear assumptions are validated, we can achieve our objectives that is predicted fitting model with higher accuracy. These four assumptions are:

- Linearity (L)
- Independent error (I)
- Normality (N)
- Equal variance (E)

From residual by predicted for registered in figure 21, three assumption of linear regression can be inferred. Relationship between residual indicate somehow a curve pattern which violate the linearity assumption. Also some cluster can be detected which disobey independent assumption. Furthermore, the variance of residuals increases from left to right that indicated the heteroscedasticity (E) which violate the equal variance. The distribution and the Q-Q plot of residual in figure 19 and 20 implies a variability of residuals. Besides, a histogram shows a right skewness which confirms the violation of normality(N).

Studentized residuals in figure 22 indicate the estimation of the standard deviation of residuals and is helpful in assessing the equal variance assumption and outliers. If 5% of all observations located over 2 or any observations located over 3, it cause a concern that our model might be poorly fitted to the data. In this plot in figure 22 number of observations over 2 (~18) is less than 5% of all cases(18/731=2.5%), and we do not have observations over 3.

Observed by predicted for the registered plot in figure 23 indicates that we have more registered variation compared to the predicted value of registered users. If the model was good, all points in the scatter plot should be along the diagonal line.

Figure 24 indicate leverage and outliers. Leverage are observations whose x-value is far away from other data points while y-value follows the regression model line. This figure shows some observations that are greater than cause of concern line and can be considered as leverage.

Figure 25 indicate Cook's Distance which is useful for identifying outliers. It shows the influence of each case on model. This figure shows no value greater than the cause of concern line(1) while some values are greater than the rule of thumb(4/count of observations). In figure 26, DFFITS also shows how influential the observation is. It indicates that some observations are greater than the rule of thumb but not greater than the cause of concern(1). In figure 27, DFBETAS measures the difference between parameter estimate with and without the influential point. It also indicates that some observations are greater than the rule of thumb but not greater than the cause of concern(1). Almost in all of them, we fit regression model with all observations, then we delete one observation and refit the regression model on the remaining. Then, we compare the results using all observations to the results with the deleted observation.

A correlation matrix relating the residuals to variables temp, hum and windspeed.

Pearson Correlation Coefficients, N = 731
Prob > |r| under H0: Rho=0

	residual_r	temp	hum	windspeed
residual_r Residual	1.00000	0.00040 0.9914	-0.19938 <.0001	-0.14007 0.0001
temp	0.00040 0.9914	1.00000	0.12696 0.0006	-0.15794 <.0001
hum	-0.19938 <.0001	0.12696 0.0006	1.00000	-0.24849 <.0001
windspeed	-0.14007 0.0001	-0.15794 <.0001	-0.24849 <.0001	1.00000

Figure 28

Number in Model	R-Square	Adjusted R-Square	C(p)	Variables in Model
1	0.2961	0.2952	60.7759	atemp
1	0.2916	0.2906	65.8480	temp
1	0.0473	0.0460	339.3082	windspeed
2	0.3247	0.3228	30.8352	atemp hum
2	0.3175	0.3156	38.8535	temp hum
2	0.3104	0.3085	46.7810	atemp windspeed
3	0.3511	0.3484	3.2708	atemp hum windspeed
3	0.3486	0.3459	6.0912	temp hum windspeed
3	0.3249	0.3221	32.6406	atemp temp hum
4	0.3513	0.3478	5.0000	atemp temp hum windspeed

Figure 29

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	624369997	156092499	98.31	<.0001
Error	726	1152741975	1587799		
Corrected Total	730	1777111972			

Root MSE	1260.07891	R-Square	0.3513
Dependent Mean	3656.17237	Adj R-Sq	0.3478
Coeff Var	34.46443		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	3326.71419	314.83365	10.57	<.0001	0
atemp	1	4013.82093	2282.92236	1.76	0.0791	63.63235
temp	1	1052.14293	2021.76391	0.52	0.6029	62.96982
hum	1	-2282.63447	340.17466	-6.71	<.0001	1.07927
windspeed	1	-3477.82005	638.79848	-5.44	<.0001	1.12677

Figure 25

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	623939982	207979994	131.12	<.0001
Error	727	1153171990	1586206		
Corrected Total	730	1777111972			

Root MSE	1259.44683	R-Square	0.3511
Dependent Mean	3656.17237	Adj R-Sq	0.3484
Coeff Var	34.44714		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	3283.66352	303.61865	10.82	<.0001	0
atemp	1	5192.07561	292.43323	17.75	<.0001	1.04517
hum	1	-2291.78924	339.54909	-6.75	<.0001	1.07638
windspeed	1	-3419.52346	628.58423	-5.44	<.0001	1.09212

Figure 26

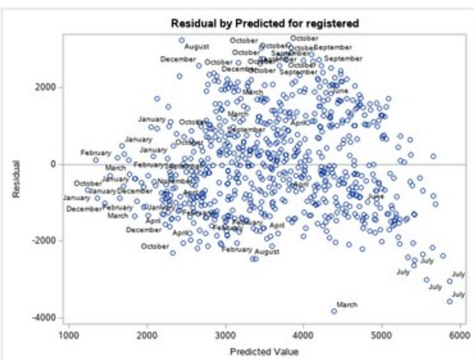


Figure 28

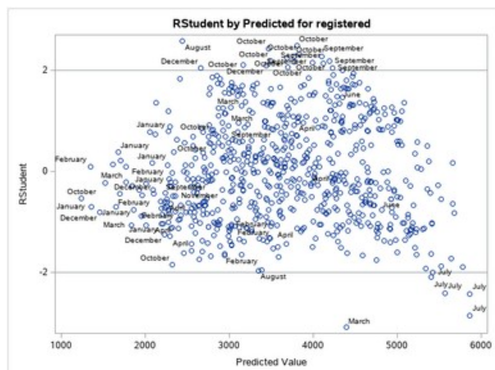


Figure 27

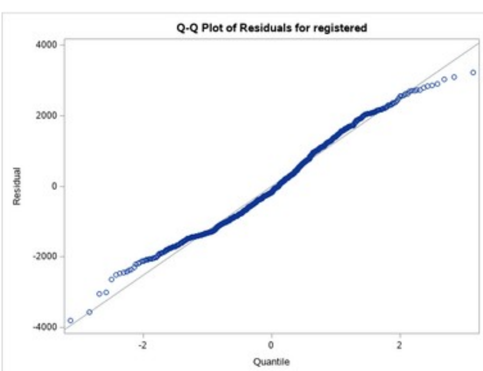


Figure 30

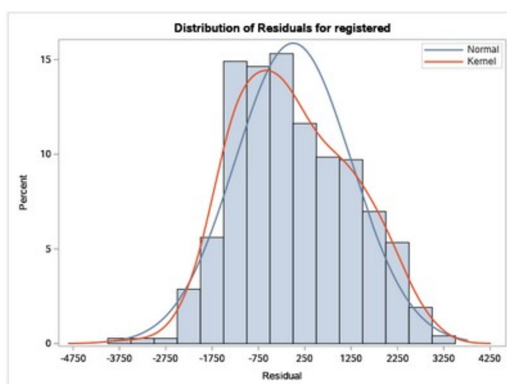


Figure 29

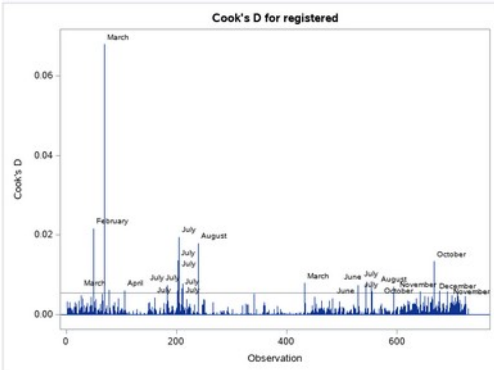


Figure 32

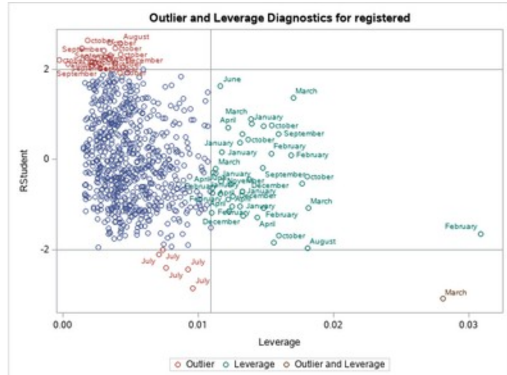


Figure 31

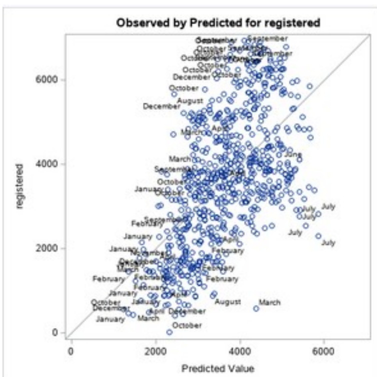


Figure 34

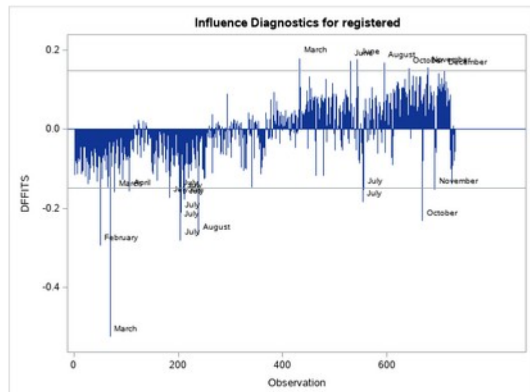


Figure 33

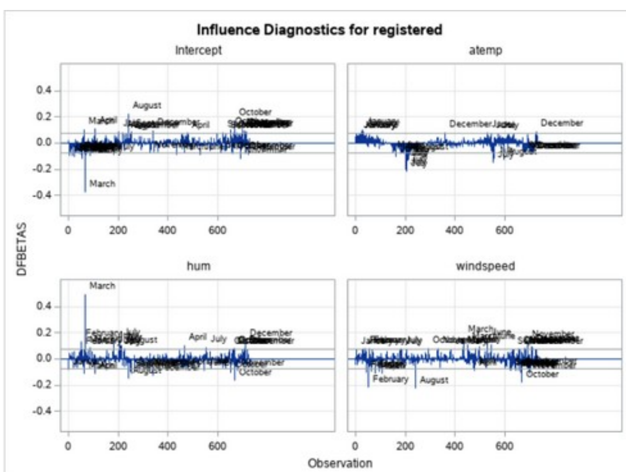


Figure 35

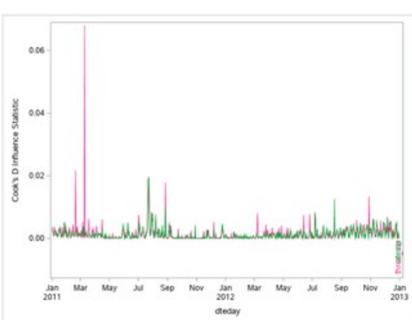


Figure 38

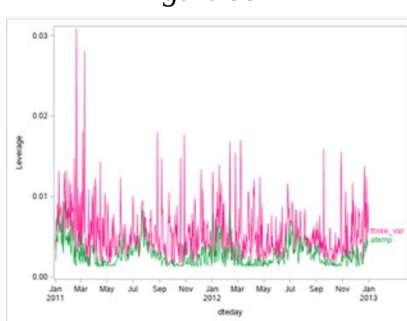


Figure 37

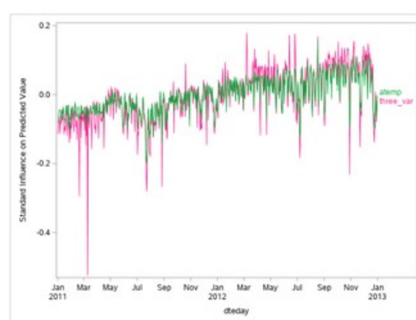


Figure 36

Observations with large residuals

Obs	registered	month	weekday	season	dteday
69	577	March	Thursday	winter	10MAR2011
203	2825	July	Friday	summer	22JUL2011
204	2298	July	Saturday	summer	23JUL2011
205	2556	July	Sunday	summer	24JUL2011
211	2916	July	Saturday	summer	30JUL2011
212	2778	July	Sunday	summer	31JUL2011
595	5665	August	Friday	summer	17AUG2012
628	6803	September	Wednesday	summer	19SEP2012
629	6781	September	Thursday	summer	20SEP2012
630	6917	September	Friday	summer	21SEP2012
634	6693	September	Tuesday	fall	25SEP2012
635	6946	September	Wednesday	fall	26SEP2012
648	5791	October	Tuesday	fall	09OCT2012
649	6911	October	Wednesday	fall	10OCT2012
650	6736	October	Thursday	fall	11OCT2012
651	6222	October	Friday	fall	12OCT2012
655	6612	October	Tuesday	fall	16OCT2012
656	6482	October	Wednesday	fall	17OCT2012
657	6501	October	Thursday	fall	18OCT2012
664	6484	October	Thursday	fall	25OCT2012
665	6262	October	Friday	fall	26OCT2012
704	6055	December	Tuesday	fall	04DEC2012
711	5219	December	Tuesday	fall	11DEC2012

Figure 39

Figure 28 indicates the correlation between residuals and temp is negligible (close to zero) and with the p-value of 0.99, we can not reject the null hypothesis (h0= there is no correlation between residual and temp). Correlation coefficient between residual with hum and windspeed implies a small effect and with p-value smaller than 0.0001 is statistically significant.

We can use a R-square method to select the best variables for our model. This method lists all possible models and we should find the highest R-square where Mallow's cp (CP) is less than or equal to p+1. In figure 29 in two rows CP condition is satisfied and can propose two models. First, model with three variables including atemp, hum and windspeed and second model with all four potential variables including atemp, temp, hum and windspeed. Both models have approximately same R-square of 0.35. also adjusted R-square which indicate how well the model we can apply to the population is near R-square. It is recommended to choose model with less variables when R-square and CP are close to each other.

In figure 30 we fit a model with maximum variables and we can see a considerable increase in R-square(0.35) compare to model with only atemp as independent variable(0.3). In other to avoid multicolinearity which indicate how two variables are tightly correlated, we can refer to variance inflation factor (VIF). Multicolinearity cause a wide variation of the estimated coefficient which leads to inaccurate model. Figure 30 shows the VIF for atemp and temp is way more than 10 (cause of concern) which implies the multicolinearity. To solve this problem we can remove one of the variables from model. Figure 31 shows after we remove temp, the F-value which indicate how well the model is, get improved (98 to 131), all parameters have a reasonable VIF and statistically significant(p-value < 0.0001) .

As residual plot in figure 33 shows the cluster point of residuals implies the issue of independent error has not yet addressed. Furthermore, the curve pattern violate the linearity assumption. The increases from left to right, indicated the heteroscedasticity which violate a equal variance. Q-Q plot and histogram of residual indicate a skewness to the right (violate normality assumption). All of these are not that bad which need to apply transformation.(log ans sqrt transformer was applied and the result became worse)

Studentized residuals in figure32 indicate no observation greater than cause of concern(1), also If 5% of all observations located over 2 it cause a concern. According table of observation with large residuals in figure 44, about 24 of our observations have a value greater than 2 and this number mean 3.2% of our data are over line 2.

Figure 36 shows some observation greater than rule of thumb which can be considered as leverage. Figure 37 indicates Cook's Distance which is useful for identifying outliers. This figure shows no value greater than the cause of concern line(1) while some values are greater than the rule of thumb. In figure 38, DFFITS indicates that some observations are greater than the rule of thumb but not greater than the cause of concern(1). In figure 27, DFBETAS indicates that some observations are greater than the rule of thumb but not greater than the cause of concern(1).

We compare the measure of influence of last model with one independent variable and recent one with three independent variables and we notice that more variable may make a model better in terms of R-square but may make some observation become more influential. In figure 43, the Cook's D indicate the model with three variables(red) have a higher influence than model with one variable. Also comparing leverage implying the increase in influence statistic of model with 3 variables, but it is noteworthy that guideline in leverage depends on predictors and in model with 3 predictor goes up($2*(p + 1)/n$). we can conclude that can not expect an improvement in observation influence just by adding independent variables to our model.

Extend multiple regression model for registered by including the variables identified in last part and categorical predictors. Considering as many potential explanatory variables as possible

Root MSE	656.13377	R-Square	0.8292
Dependent Mean	3656.17237	Adj R-Sq	0.8232
Coeff Var	17.94592		

Figure 40

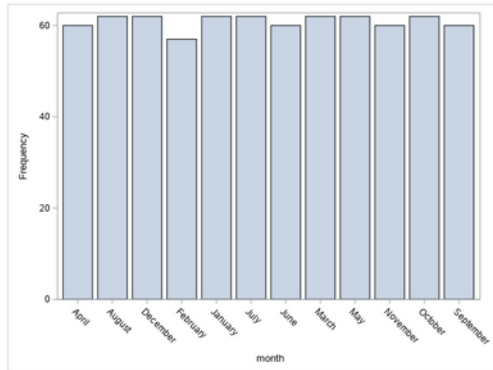


Figure 41

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	25	1473601350	58944054	136.92	<.0001
Error	705	303510622	430512		
Corrected Total	730	1777111972			

Figure 42

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2007.83669	288.28908	6.96	<.0001
atemp	1	3628.65027	358.89574	10.11	<.0001
hum	1	-2304.75434	191.56489	-12.03	<.0001
windspeed	1	-2422.45595	336.93440	-7.19	<.0001
workingday	1	1048.36330	153.32747	6.84	<.0001
yr	1	1726.57149	49.35388	34.98	<.0001
February	1	114.21929	122.54802	0.93	0.3516
March	1	206.47858	139.99659	1.47	0.1407
April	1	112.50182	209.80350	0.54	0.5920
May	1	464.63705	223.98716	2.07	0.0384
June	1	296.03514	232.28006	1.27	0.2029
July	1	-163.18234	259.74316	-0.63	0.5300
August	1	252.50876	248.96236	1.01	0.3108
September	1	625.79959	221.34490	2.83	0.0048
October	1	33.58115	204.65328	0.16	0.8697
November	1	-313.36669	196.57222	-1.59	0.1113
December	1	-130.00503	155.34316	-0.84	0.4029
summer	1	74.22518	157.42602	0.47	0.6374
fall	1	879.19310	180.99953	4.86	<.0001
winter	1	-636.85927	153.03402	-4.16	<.0001
Saturday	1	262.86196	175.76683	1.50	0.1352
Sunday	1	32.15130	175.79501	0.18	0.8549
Monday	1	-102.93615	92.94036	-1.11	0.2684
Tuesday	1	24.77242	91.33781	0.27	0.7863
Wednesday	1	77.18067	91.29702	0.85	0.3982
Thursday	1	85.29059	91.20467	0.94	0.3500

Figure 43

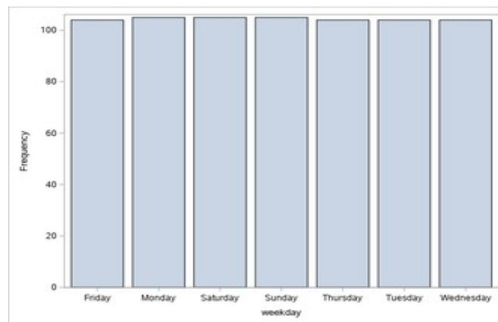


Figure 44

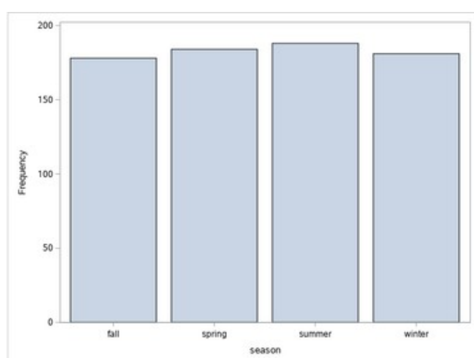


Figure 45

Stepwise Selection: Step 14

Variable December Entered: R-Square = 0.8277 and C(p) = 10.0755

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	14	1470985759	105070411	245.75	<.0001
Error	716	306126214	427551		
Corrected Total	730	1777111972			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	2097.17091	211.52671	42026644	98.30	<.0001
atemp	4039.79761	260.99148	102436336	239.59	<.0001
hum	-2359.74493	186.14675	68707904	160.70	<.0001
windspeed	-2459.07815	331.44336	23534962	55.05	<.0001
workingday	1065.10350	65.48136	113119045	264.57	<.0001
yr	1718.53895	48.95679	526842074	1232.23	<.0001
May	222.56063	95.49308	2322418	5.43	0.0200
July	-395.08827	102.13337	6397944	14.96	0.0001
September	470.23270	95.16716	10438520	24.41	<.0001
November	-336.79746	112.65184	3821629	8.94	0.0029
December	-167.07776	99.88962	1196146	2.80	0.0948
fall	738.71327	85.64406	31808617	74.40	<.0001
winter	-687.57517	92.82143	23460182	54.87	<.0001
Saturday	234.87807	86.89723	3123637	7.31	0.0070
Monday	-150.99565	70.14027	1981442	4.63	0.0317

Figure 46

Root MSE	653.87352	R-Square	0.8277
Dependent Mean	3656.17237	Adj R-Sq	0.8244
Coeff Var	17.88410		

Figure 47

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	2097.17091	211.52671	9.91	<.0001	0
atemp	1	4039.79761	260.99148	15.48	<.0001	3.08856
hum	1	-2359.74493	186.14675	-12.68	<.0001	1.20017
windspeed	1	-2459.07815	331.44336	-7.42	<.0001	1.12651
workingday	1	1065.10350	65.48136	16.27	<.0001	1.58457
yr	1	1718.53895	48.95679	35.10	<.0001	1.02446
May	1	222.56063	95.49308	2.33	0.0200	1.21020
July	1	-395.08827	102.13337	-3.87	0.0001	1.38436
September	1	470.23270	95.16716	4.94	<.0001	1.16666
November	1	-336.79746	112.65184	-2.99	0.0029	1.63473
December	1	-167.07776	99.88962	-1.67	0.0948	1.32420
fall	1	738.71327	85.64406	8.63	<.0001	2.31012
winter	1	-687.57517	92.82143	-7.41	<.0001	2.74431
Saturday	1	234.87807	86.89723	2.70	0.0070	1.58807
Monday	1	-150.99565	70.14027	-2.15	0.0317	1.03465

Figure 48

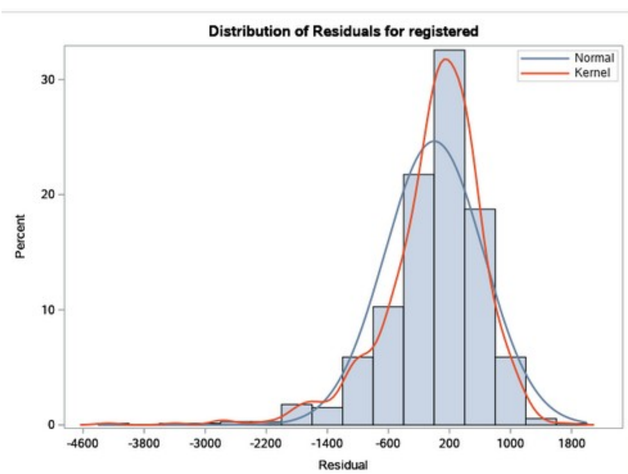


Figure 50

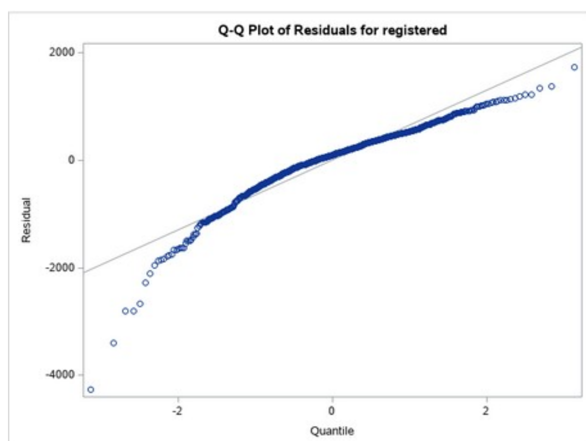


Figure 49

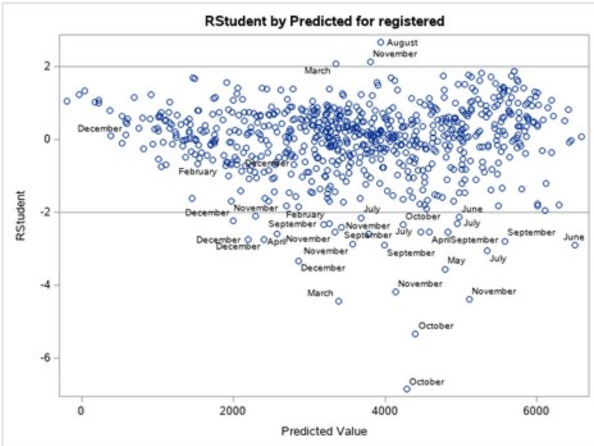


Figure 51

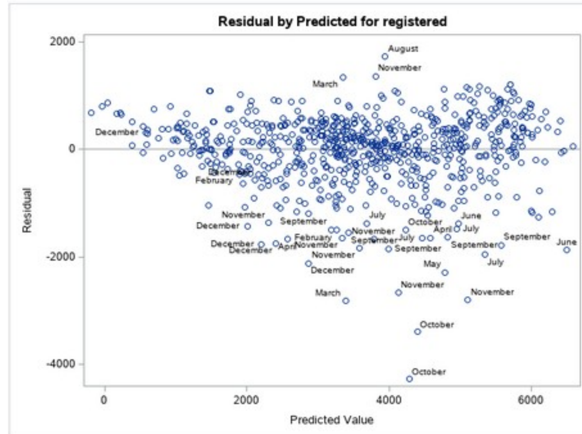


Figure 52

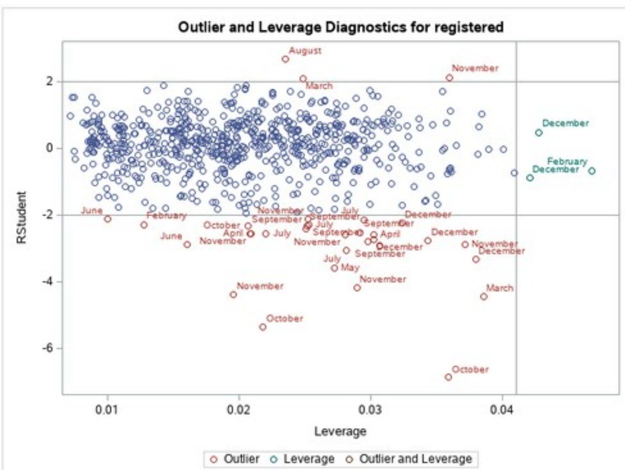


Figure 54

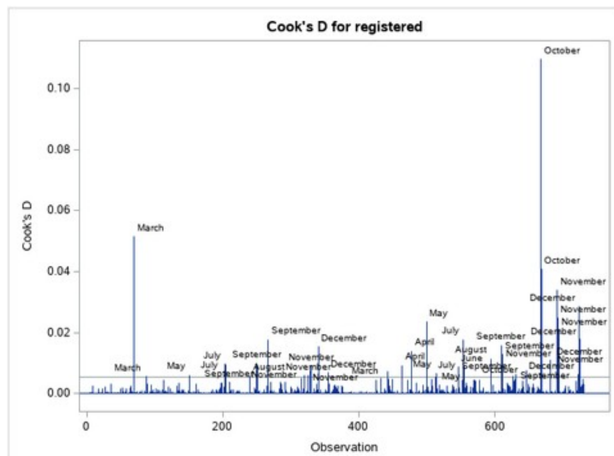


Figure 53

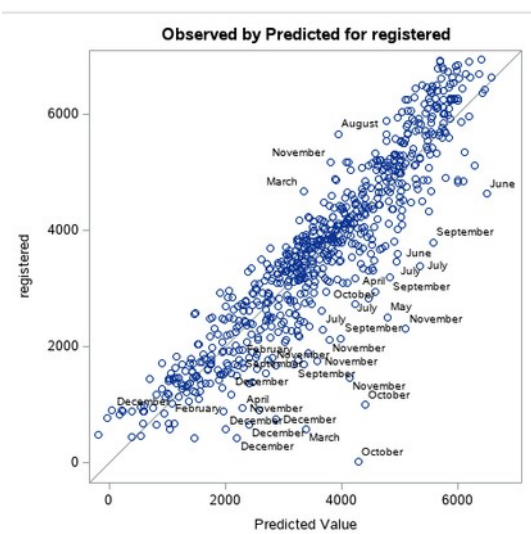


Figure 55

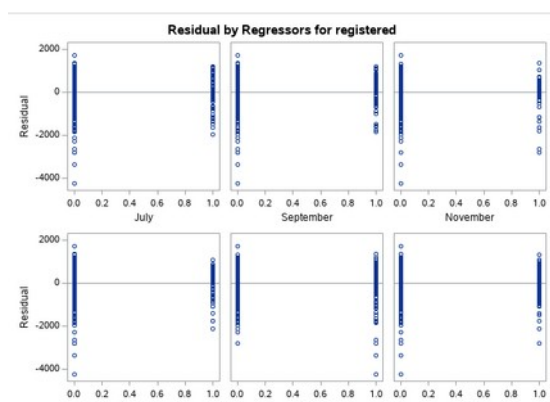


Figure 56

Obs	registered	month	weekday
69	577	March	Thursday
203	2825	July	Friday
204	2298	July	Saturday
250	1878	September	Wednesday
251	1689	September	Thursday
266	2137	September	Friday
328	935	November	Thursday
329	1697	November	Friday
341	655	December	Wednesday
425	1769	February	Wednesday
442	4681	March	Saturday
464	2939	April	Sunday
478	907	April	Sunday
500	2501	May	Monday
518	3594	June	Friday
546	4634	June	Friday
554	3392	July	Saturday
555	3469	July	Sunday
595	5665	August	Friday
610	3788	September	Saturday
611	3197	September	Sunday
646	2729	October	Sunday
668	20	October	Monday
669	1009	October	Tuesday
682	5172	November	Monday
692	1470	November	Thursday
693	2307	November	Friday
694	1745	November	Saturday
695	2115	November	Sunday
724	746	December	Monday
725	573	December	Tuesday
726	432	December	Wednesday

Figure 58- large residuals

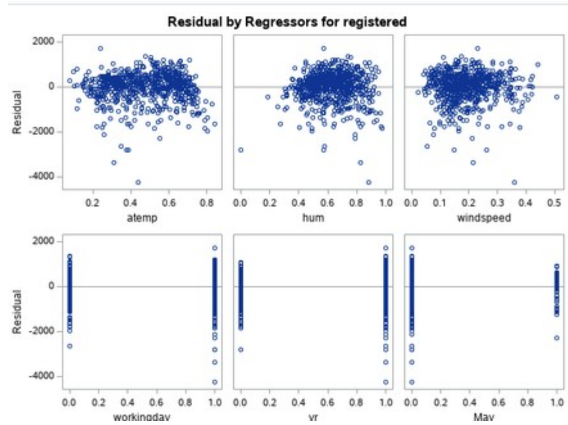


Figure 57

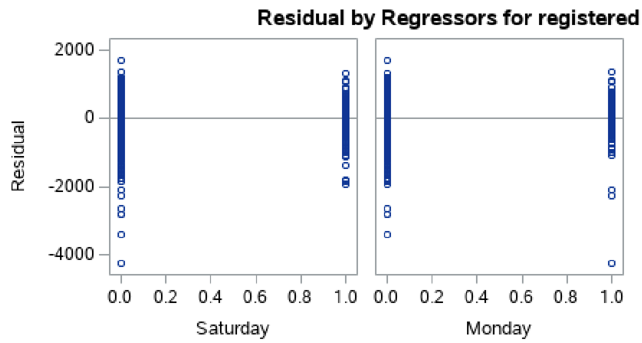


Figure 59

We can extend our regression model by including year, working day. Also to user categorical variables in our model such as month, weekday and season we need to use dummy method. The variable with lower frequency is suitable as a baseline. In order to define dummy variables we consider January, Friday and Spring as a baseline for month, weekday and season respectively. Figure 46,49 and 50 show the frequency of categorical variables. The frequency in all groups of categorical variable approximately are same and lets us to choose each group to remove as a baseline.

In figure 48 we have all 25 possible variables except for temp, casual and count as independent variables of our regression model. As we see in figure 45, R-square increase dramatically from about 0.35 (last model with three variables) to 0.83. It means variables in this model account for about 83% of variability in registered users and imply a strong association with response variable. This model has F-value equal to about 137 which indicate a good model and is statistically significant (p-value <0.0001). There are some variables in this model which are not statistically significant.

In order to avoid overfitting and have a simpler model with lower number of variables, we use stepwise selection method. This model starts with simple linear regression model (with one variable) and increase the variable to reach the highest R-square. After 14 step it reaches to 14 variables which all variables are statistically significant at 0.15 level(other variables that are not exist in this model, are not statistically significant). According to figure 51, F-value increases significantly compare to model with 25 variables. A good point about this simpler model is, Although we decrease the independent variables to 14, the R-square has not decreased significantly and stayed at about 0.83. Also adjusted R-square is closely behind as well, at about 0.82 and imply the quality of model is very good to the population. This model is a BEST model without applying any filtration on outliers.

According to figure 51, as regression formula implies, with each degree increase in atemp while other variable be same and do not change, the number of registered increase by 4039 users. About humidity and wind speed, this is inverse and by each unit increase registered users decreases by 2359 and 2459 respectively. About year variable in dataset, number 1 refers to 2012 and 0 refers to 2011, we can say when everything else are same, the registered user in 2012 are 1718 users higher than 2011. About categorical variables like season we compare variable with baseline, for instance, in July the number of registered user decrease by 395 compare to January. In Winter and Monday the number of registered user decrease by 687 and 150 respectively compare to Spring and Friday if everything else is the same.

In order to check assumption for linear regression first we need to check variance inflation factor to avoid multicolinearity. As figure 53 shows all VIF for our variables are acceptable and below 10. The residual plot in figure 52 shows no pattern that imply violation of linearity and independent error but the variance of residual points are not yet equal. Figure 56, 57 and 59 shows an unequal variance of residual in roughly all independent variables. The histogram of residual is left-skewed which some outliers on left are evident. Furthermore, the Q-Q plot implies the negative skewness.

Figure 60 shows that actual registered and predicted registered follow the diagonal line. If we look carefully at influence of observation in cook’s D, we can see a bunch of point which above rule of thumb, but is not above 1 and are not cause of concern. Rstudent shows outliers, according to figure 61, there are 32 observations with large residuals and this number account for less than 5 percent of observation(32/731=0.04). Also according to figure 59, there are three observation considered as a leverage with values greater than rule of thumb.

We consider the independent variable somehow as a cause while dependent variable as effect. The dependent variable as its name suggest depends on independent variables. We can choose registered, count or casual as dependent variable because they are dependent on other variable and should not be used as explanatory. To be precise the change in season, month, weekdays and weather can affect count or casual variable and use them as explanatory variables is not reasonable. On the other hand, count variable is summation of casual and registered, and it is better not to use derived variable which might cause multicolinearity.

Model 1: Applying rule of thumb as filtration

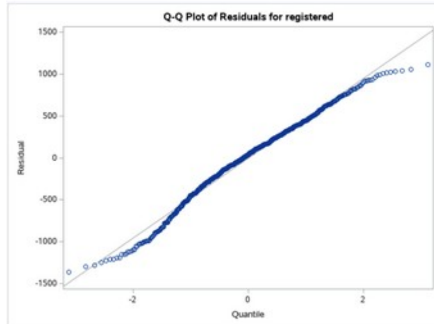
Dependent Variable: registered

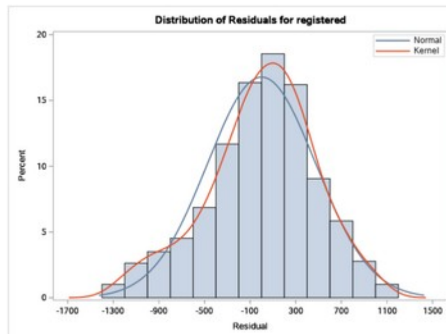
Number of Observations Read	685
Number of Observations Used	685

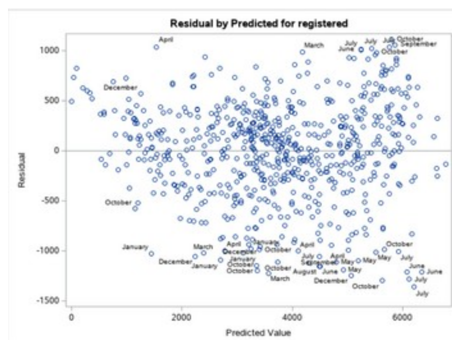
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	14	1441728335	102980595	445.94	<.0001
Error	670	154723672	230931		
Corrected Total	684	1596452007			

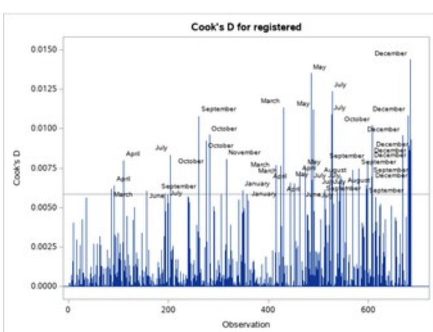
Root MSE	480.55265	R-Square	0.9031
Dependent Mean	3746.07007	Adj R-Sq	0.9011
Coeff Var	12.82818		

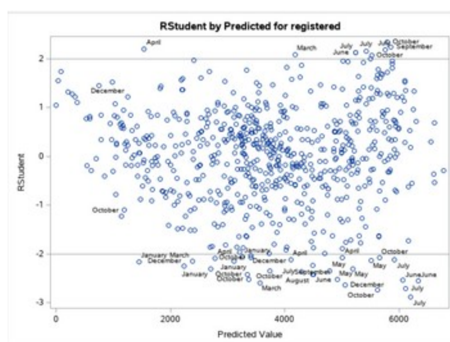
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1939.73133	164.98009	11.76	<.0001
atemp	1	4122.37576	201.75315	20.43	<.0001
hum	1	-2156.92652	148.26016	-14.55	<.0001
windspeed	1	-2169.64755	258.17397	-8.40	<.0001
workingday	1	991.48679	50.77037	19.53	<.0001
yr	1	1786.90715	37.34249	47.85	<.0001
May	1	268.99362	72.38370	3.72	0.0002
July	1	-329.60276	76.74866	-4.29	<.0001
September	1	560.48373	73.54120	7.62	<.0001
November	1	-143.06743	88.61094	-1.61	0.1069
December	1	-86.98476	79.77851	-1.09	0.2760
fall	1	774.06790	65.15173	11.88	<.0001
winter	1	-662.59472	71.59594	-9.25	<.0001
Saturday	1	156.80632	67.18616	2.33	0.0199
Monday	1	-143.93530	53.38401	-2.70	0.0072

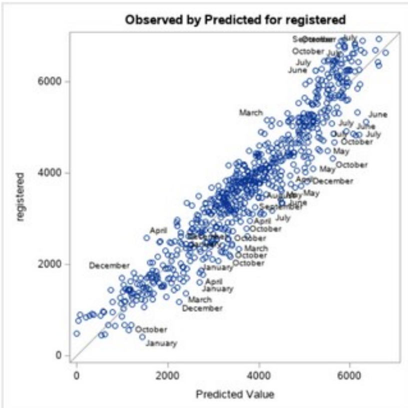


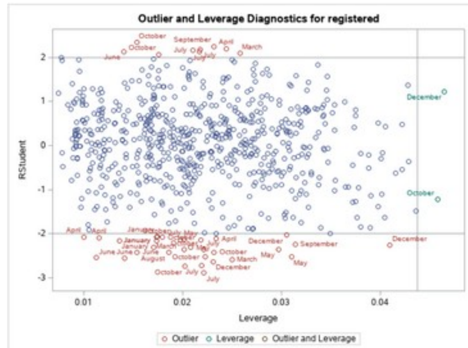












We apply filter to get a model with less noise and higher R-square and F-value. By applying rule of thumb to filter outlier, leverage and influential observation, the number of observation decrease from 731 to 685 and we miss some part of out data. Observation which their Rstudent more than 2, their leverage more than 0.041 (30/731), Cook D more than 0.0054 (4/731) and Dffit more than 0.405 (2*sqrt(30/731)) was removed. This intersection of rules help us to remove observation with maximum influence. In the new model, variables including December and November are not statistically significant at level of 0.05. The F-value increase dramatically from 245 (in BEST model) to 446 which indicates how our model get improved. Furthermore, R-square increases and reach to about 0.9 which implies an excellent association between independent variables and dependent.

After filtering, the histogram of residual depict a symmetric shape and Q-Q plot with negligible variation at both sides confirm a normality(N). As residual plots show the condition for confirming independent error(I) and linearity(L) get improved and an equal variance is evident.

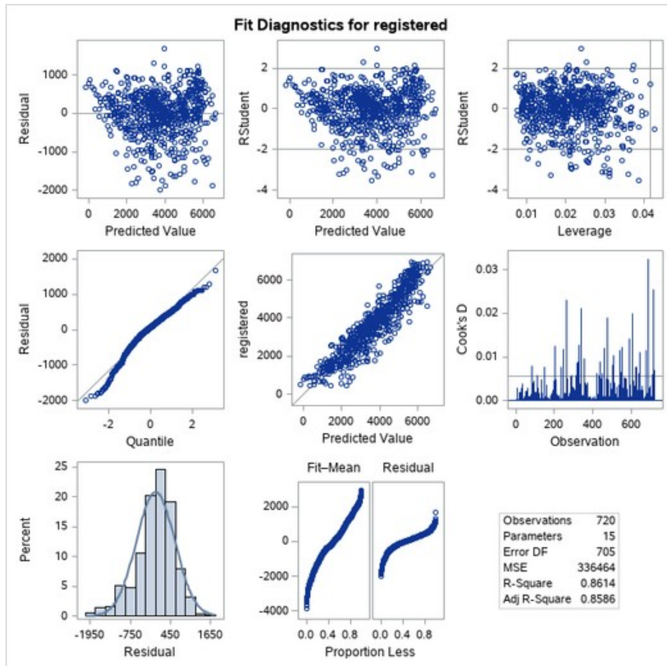
Model 2: Applying cause of concern as filtration

Number of Observations Read	720
Number of Observations Used	720

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	14	1474228075	105302005	312.97	<.0001
Error	705	237207383	336464		
Corrected Total	719	1711435458			

Root MSE	580.05549	R-Square	0.8614
Dependent Mean	3690.42500	Adj R-Sq	0.8586
Coeff Var	15.71785		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2026.32697	189.41462	10.70	<.0001
atemp	1	4007.24026	233.75359	17.14	<.0001
hum	1	-2339.91151	169.85060	-13.78	<.0001
windspeed	1	-2291.55152	300.99923	-7.61	<.0001
workingday	1	1081.46911	58.54202	18.47	<.0001
yr	1	1768.62052	43.91484	40.27	<.0001
May	1	259.01905	85.27211	3.04	0.0025
July	1	-355.87776	91.04172	-3.91	0.0001
September	1	451.80269	84.55350	5.34	<.0001
November	1	-322.62077	101.72887	-3.17	0.0016
December	1	-189.46616	91.50298	-2.07	0.0388
fall	1	811.93547	76.81161	10.57	<.0001
winter	1	-666.96345	83.18299	-8.02	<.0001
Saturday	1	254.07742	77.94013	3.26	0.0012
Monday	1	-92.48129	63.29243	-1.46	0.1444



If we apply cause of concern to filter outlier, leverage and influential observation, the number of removed observation decrease to just 11 items(rule of thumb removed 44 items). As can be seen the F-value is equal 312 which is statistically significant and R-square is 0.86 which are higher than model without any filtering but are less than last model. Also in comparison with model without filtering the condition for accepting all linear regression assumption get improved.

A summary

In our dataset we have a number of registered users for roughly all days of two years. Our analysis implies the number of users varies by weekday, month and year. We have more applicant in summer and in 2012 and less in winter and in 2011. Outliers and minimum value in some plots indicate a sudden decrease in biking demand implying special event.

Initial we use a single regression model to predict registered users. We have single independent variable atemp which affect on registered user. The R-square in the regression model shows that independent variables account for about 30 percent variability in registered users. The model is statistically significant but in order to improve R-square we need to add other possible variables.

In order to fit a multiple regression model to predict registered users with higher accuracy, we can use other numerical variables in our dataset. Those variables with high correlation with registered and less correlation with each other can more contribute to our model. We add Hum and windspeed as independent variables and we witness an increase in the model efficiency(R-square=0.35)

Measuring regression assumptions, outlier, leverage and influential points help us to evaluate our model. In order to increase model accuracy, we can use categorical variables in our dataset by dummy method. After applying year, working day, weekday, month and season as independent variables, the number of variables in this model reaches to 25. The model accuracy improved significantly and R-square reaches to 0.83 which means these variables are account for 83% variability in our model. Also all conditions to accept regression assumptions get improved. To avoid overfitting and have a simpler model, we can apply stepwise method to decrease the number of independent variables to 14, while the model accuracy has roughly remained unchanged.

To increase the accuracy of model, we need to apply some filtering on data. By applying rule of thumb to filter outlier, leverage and influential observation, the number of observation decrease we miss some part of out data, but we get a model with higher accuracy (R-square), also the conditions to accept the assumptions get improved.