Project 2

Owen McGrattan

4/24/2017

Question 1

```
#Read in training data
library(readr)
train <- read_csv("~/stat28/projects/data/train.csv")
```

```
## Parsed with column specification:
## cols(
##   datetime = col_datetime(format = ""),
##   season = col_integer(),
##   holiday = col_integer(),
##   workingday = col_integer(),
##   weather = col_integer(),
##   temp = col_double(),
##   atemp = col_double(),
##   humidity = col_integer(),
##   windspeed = col_double(),
##   casual = col_integer(),
##   registered = col_integer(),
##   count = col_integer()
## )
```

Question 2

```
summary(train)
```

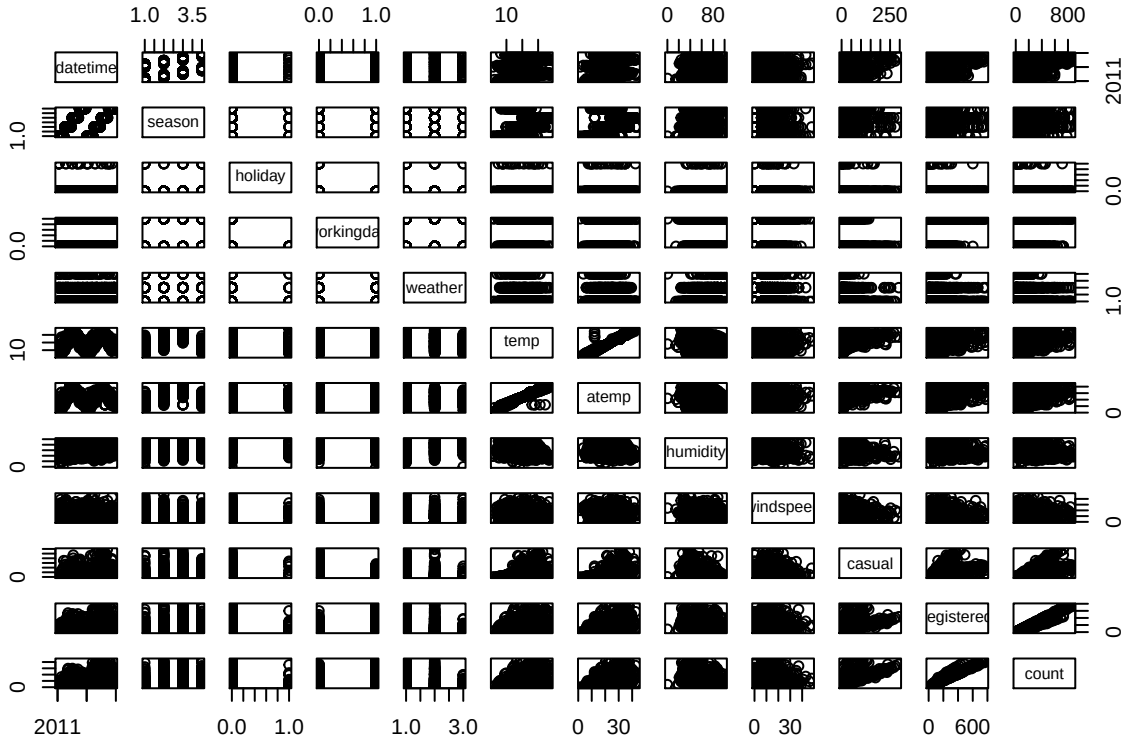
```
## Warning in as.POSIXlt.POSIXct(x, tz): unknown timezone 'zone/tz/2017c.1.0/
## zoneinfo/America/Los_Angeles'
```

```
##     datetime                       season         holiday
##  Min.   :2011-01-01 00:00:00   Min.   :1.000   Min.   :0.00000
##  1st Qu.:2011-07-02 07:15:00   1st Qu.:2.000   1st Qu.:0.00000
##  Median :2012-01-01 20:30:00   Median :3.000   Median :0.00000
##  Mean   :2011-12-27 05:56:22   Mean   :2.507   Mean   :0.02857
##  3rd Qu.:2012-07-01 12:45:00   3rd Qu.:4.000   3rd Qu.:0.00000
##  Max.   :2012-12-19 23:00:00   Max.   :4.000   Max.   :1.00000
##    workingday        weather           temp           atemp
##  Min.   :0.0000   Min.   :1.000   Min.   : 0.82   Min.   : 0.76
##  1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:13.94   1st Qu.:16.66
##  Median :1.0000   Median :1.000   Median :20.50   Median :24.24
##  Mean   :0.6809   Mean   :1.418   Mean   :20.23   Mean   :23.66
##  3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:26.24   3rd Qu.:31.06
##  Max.   :1.0000   Max.   :4.000   Max.   :41.00   Max.   :45.45
##     humidity        windspeed          casual         registered
##  Min.   :  0.00   Min.   : 0.000   Min.   :  0.00   Min.   :  0.0
##  1st Qu.: 47.00   1st Qu.: 7.002   1st Qu.:  4.00   1st Qu.: 36.0
##  Median : 62.00   Median :12.998   Median : 17.00   Median :118.0
##  Mean   : 61.89   Mean   :12.799   Mean   : 36.02   Mean   :155.6
##  3rd Qu.: 77.00   3rd Qu.:16.998   3rd Qu.: 49.00   3rd Qu.:222.0
```

```
##  Max.   :100.00   Max.   :56.997   Max.   :367.00   Max.   :886.0
##      count
##  Min.   :  1.0
##  1st Qu.: 42.0
##  Median :145.0
##  Mean   :191.6
##  3rd Qu.:284.0
##  Max.   :977.0
```

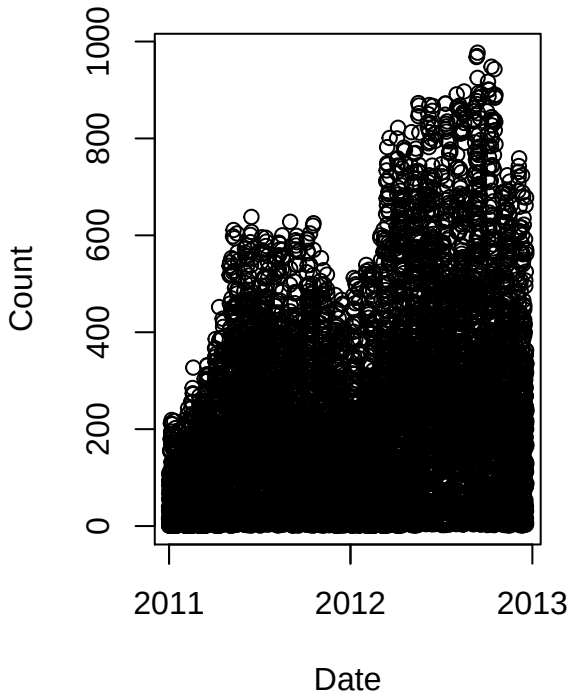
Since each row in the data frame represents an hour with each of these different variables there is going to be a lot of deviation in the count variable as well as the casual and registered variables.

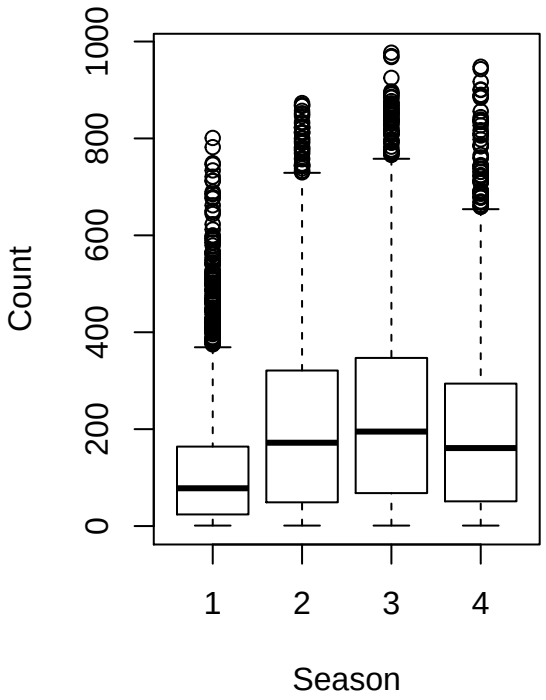
```
random<-train[sample(nrow(train),1000),]
pairs(random)
```



After looking at a pairs plot of a random sample of 1000 timestamps (was not readable with the full train dataset), there appears to be something of an increase in the count as the date increases

```
par(mfrow=c(1,2))
plot(train$datetime,train$count,xlab="Date",ylab="Count")
boxplot(count~season,data=train,ylab="Count",xlab="Season")
```

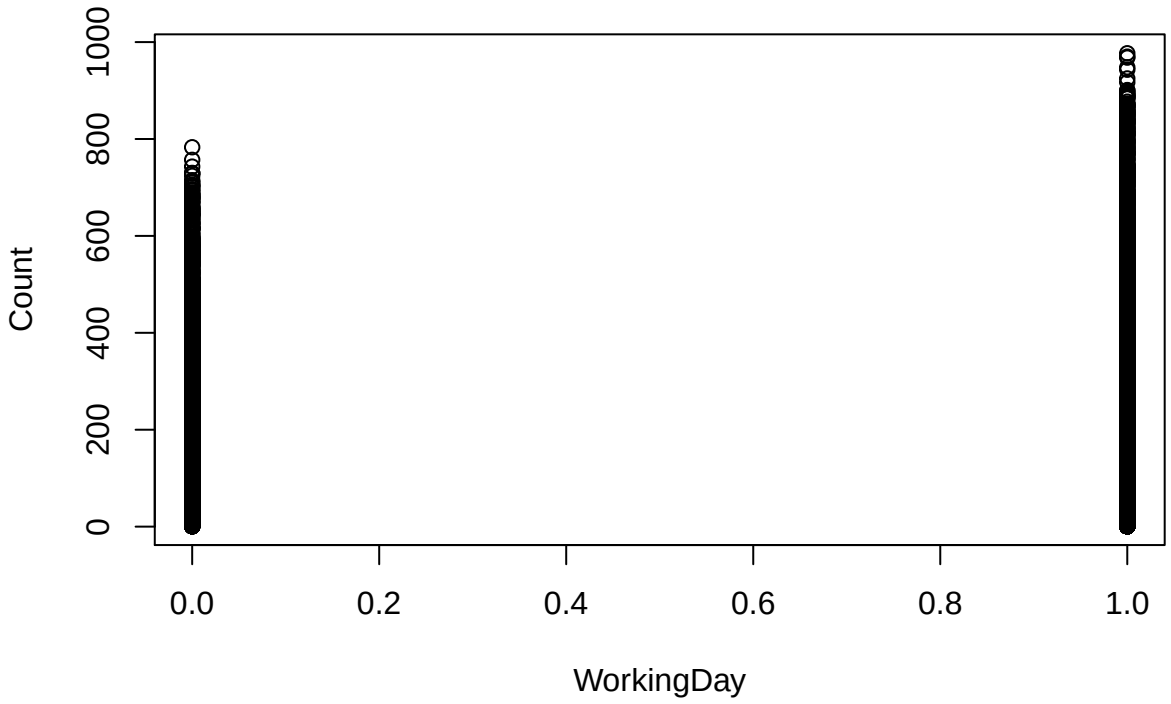




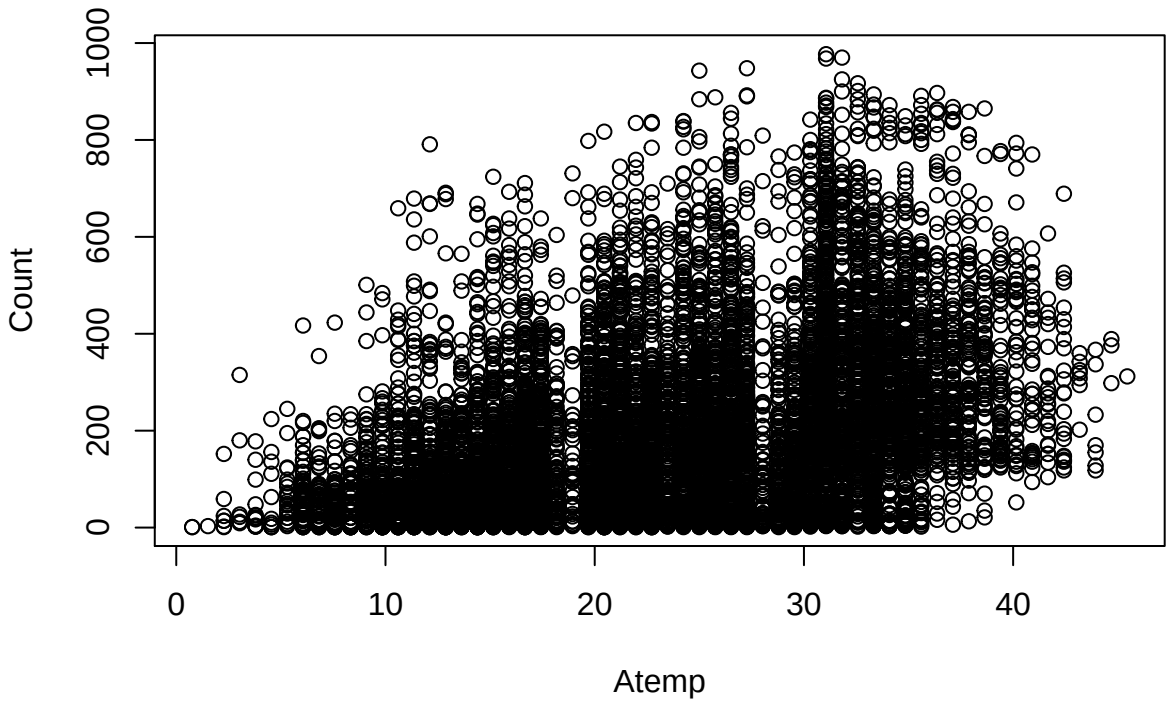
From the first plot we can note that the number of rentals each day has gone up considerably, notably in the summer and the fall when rentals appear to be their highest. Bike sharing in these cities has appeared to have grown considerably.

Question 3

```
plot(train$workingday,train$count,xlab="WorkingDay",ylab="Count")
```



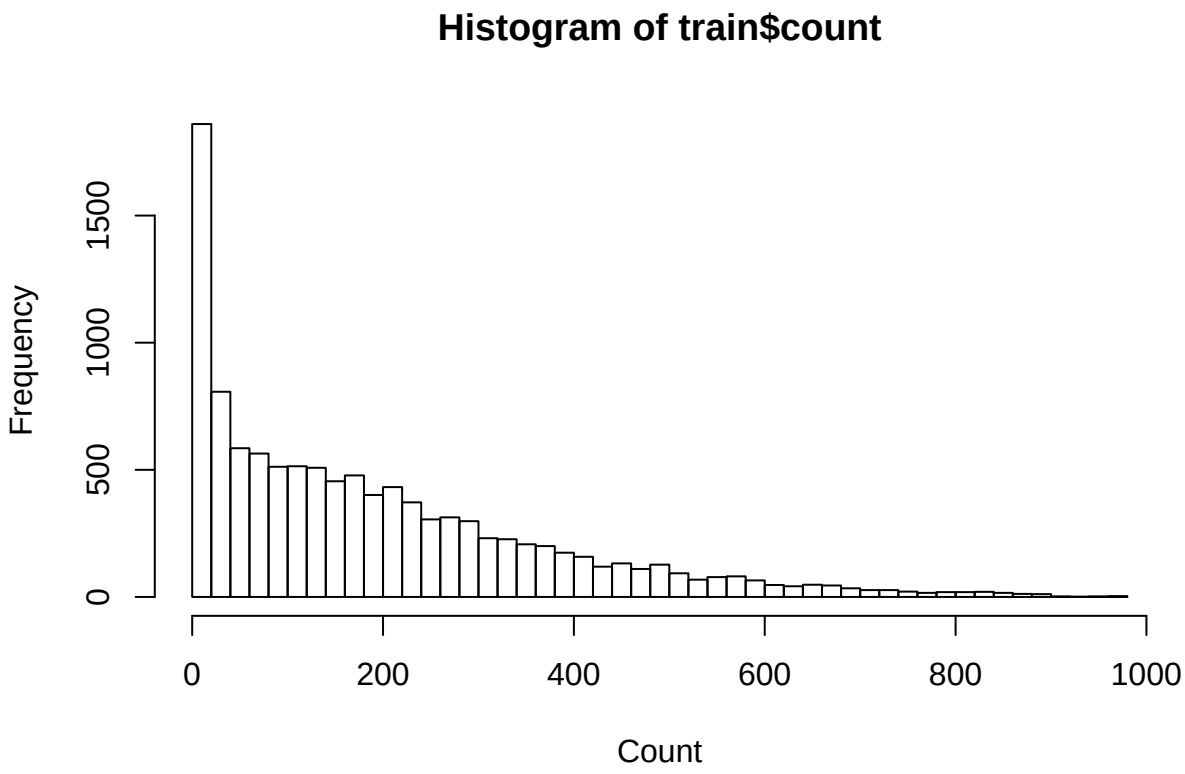
```
plot(train$atemp,train$count,xlab="Atemp",ylab="Count")
```



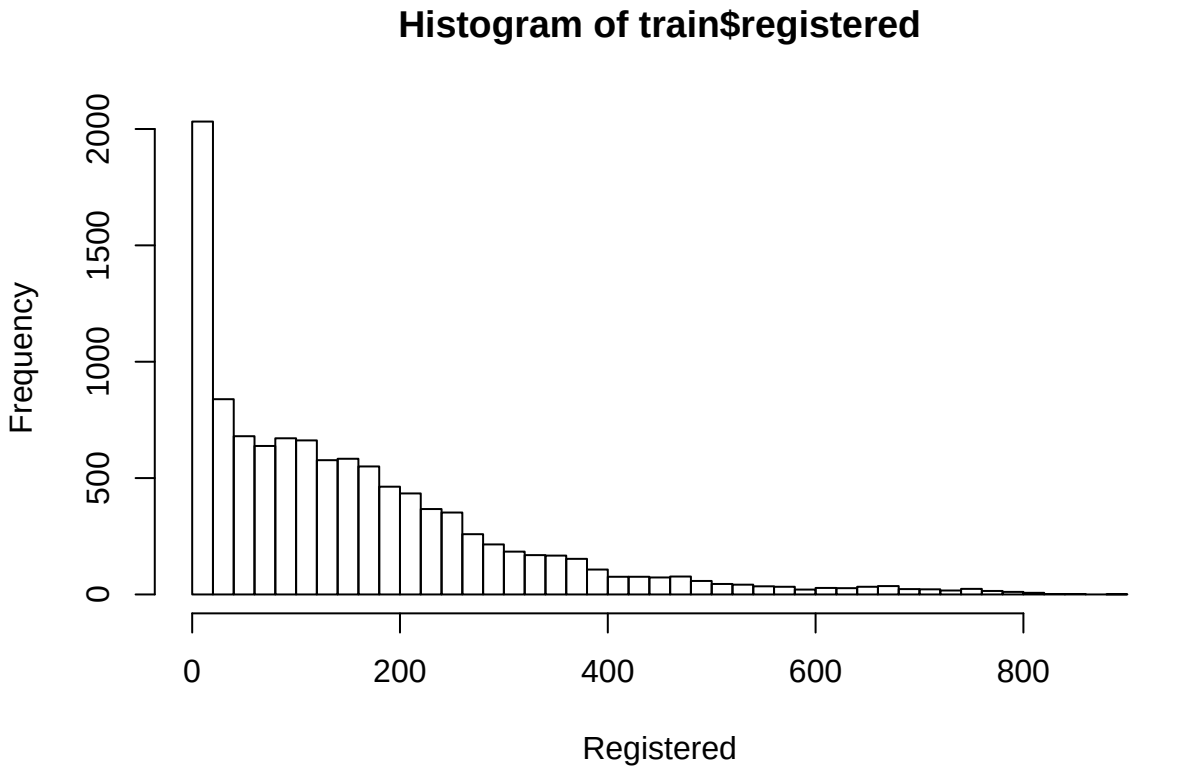

```
num.work<-nrow(train[train$workingday==1,])
num.nonwork<-nrow(train[train$workingday==0,])
```

As would be expected, there are more rentals on days when the weather is warmer but begins to tail off when it feels warmer than 40 degrees celcius outside. There appear to be higher rental counts on working days but this could be a result of the fact that there are a total of ~ 308 working days in the train dataset compared to ~ 144 non working days.

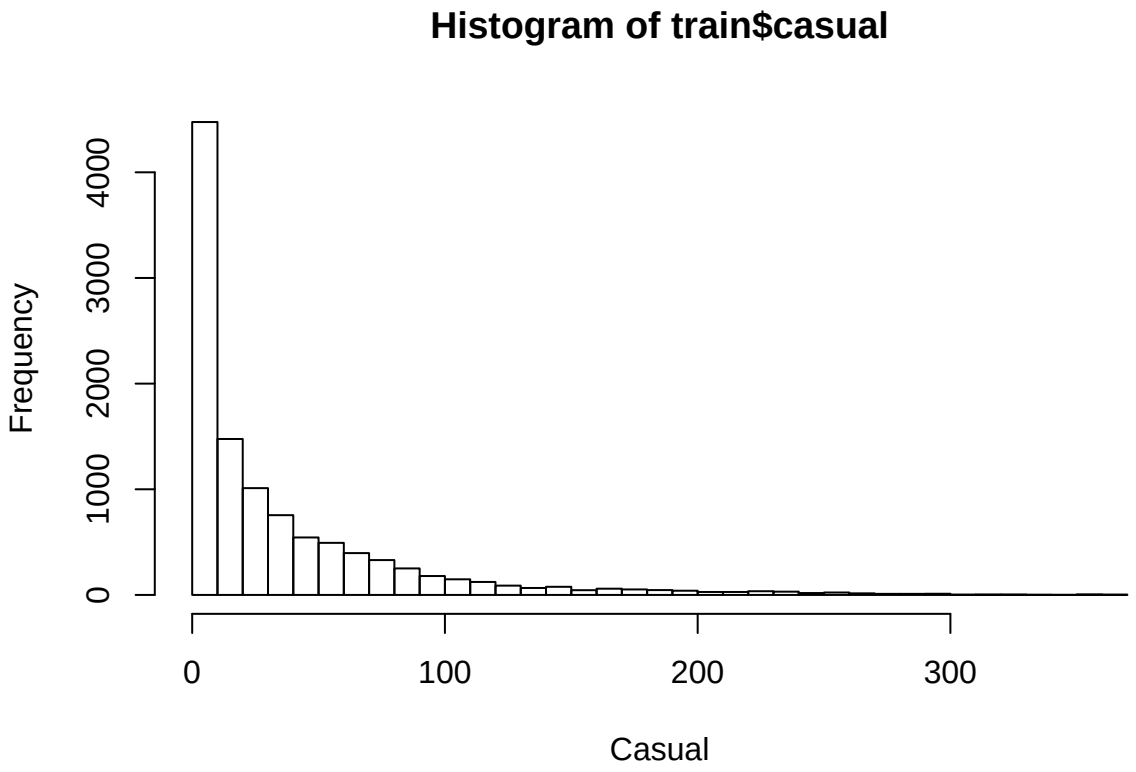
```
hist(train$count,breaks=50,xlab="Count")
```




```
hist(train$registered,breaks=50,xlab="Registered")
```

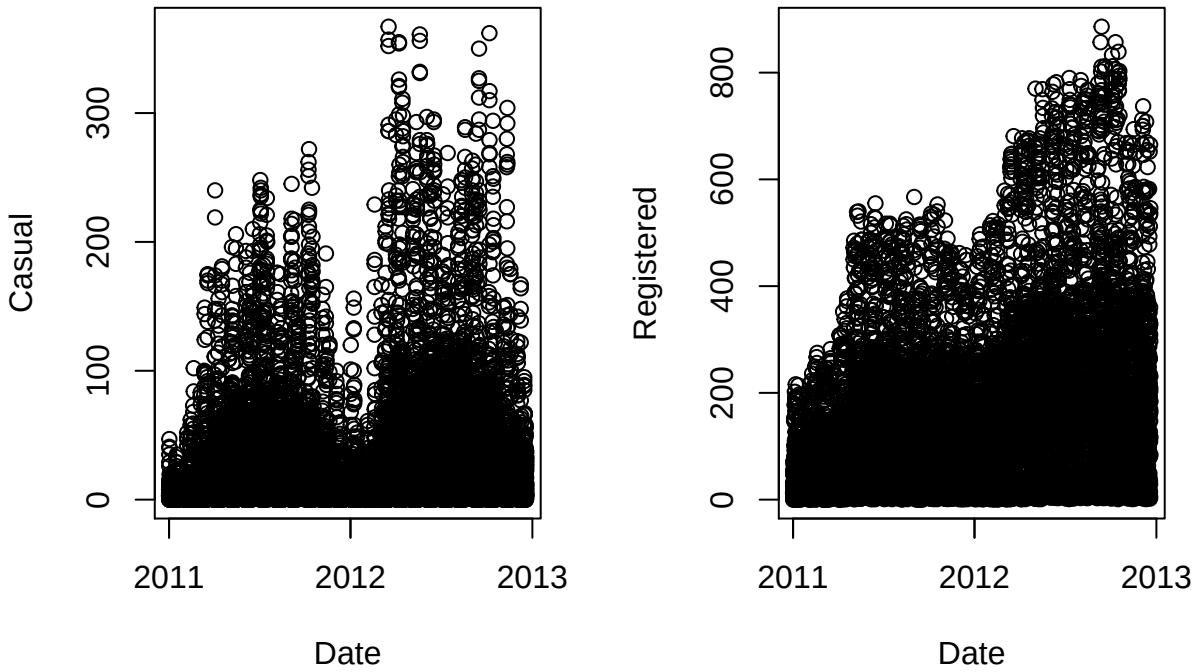



```
hist(train$casual,breaks=50,xlab="Casual")
```



Due to the heavy right skewness in the distributions of casual, registered, and count it is appropriate to do log transformations on both casual and registered.

```
par(mfrow=c(1,2))
plot(train$datetime,train$casual,xlab="Date",ylab="Casual")
plot(train$datetime,train$registered,xlab="Date",ylab="Registered")
```



Given the general differences in the number of casual and registered riders as well as the differing variations and distributions, separate models for the number of casual and registered riders should be called for with log(casual) and log(registered) as the two response variables.

Question 4

```
#Make sure there are no zero values in casual and registered when running log transformations,
#add 1 to casual and registered.

#Regression for log(casual) and log(registered)
train$lcasual<-log(train$casual+1)
train$lregistered<-log(train$registered+1)
casual_fit<-lm(lcasual~.,data=train[,-c(12,11,10,14)])
registered_fit<-lm(lregistered~.,data=train[,-c(12,10,11,13)])
summary(registered_fit)
```

```
##
## Call:
## lm(formula = lregistered ~ ., data = train[, -c(12, 10, 11, 13)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2832 -0.5744  0.2381  0.8062  3.2474
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.403e+01  9.608e-01 -14.601  < 2e-16 ***
## datetime     1.382e-08  7.327e-10  18.863  < 2e-16 ***
## season       8.981e-02  1.247e-02   7.200 6.40e-13 ***
## holiday     -5.345e-02  7.219e-02  -0.740   0.4591
## workingday   5.702e-02  2.582e-02   2.208   0.0272 *
## weather      1.115e-01  2.041e-02   5.466 4.70e-08 ***
## temp        -7.511e-04  8.896e-03  -0.084   0.9327
## atemp        4.456e-02  8.182e-03   5.446 5.25e-08 ***
## humidity    -2.382e-02  7.233e-04 -32.939  < 2e-16 ***
## windspeed    9.901e-03  1.559e-03   6.351 2.23e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.213 on 10876 degrees of freedom
## Multiple R-squared:  0.249,  Adjusted R-squared:  0.2483
## F-statistic: 400.6 on 9 and 10876 DF,  p-value: < 2.2e-16
```

```
summary(casual_fit)
```

```
##
## Call:
## lm(formula = lcasual ~ ., data = train[, -c(12, 11, 10, 14)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0485 -0.6401  0.1503  0.7548  3.1494
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.626e+00  8.365e-01  -6.726 1.83e-11 ***
## datetime     6.010e-09  6.379e-10   9.421  < 2e-16 ***
## season       9.001e-02  1.086e-02   8.288  < 2e-16 ***
## holiday     -2.619e-01  6.286e-02  -4.167 3.11e-05 ***
```

```
## workingday  -7.493e-01  2.248e-02 -33.331  < 2e-16 ***
## weather      7.250e-02  1.777e-02   4.081 4.52e-05 ***
## temp         1.803e-02  7.745e-03   2.327     0.02 *
## atemp        7.577e-02  7.124e-03  10.637  < 2e-16 ***
## humidity    -2.758e-02  6.297e-04 -43.795  < 2e-16 ***
## windspeed    8.993e-03  1.357e-03   6.625 3.63e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.056 on 10876 degrees of freedom
## Multiple R-squared:  0.4985, Adjusted R-squared:  0.4981
## F-statistic:  1201 on 9 and 10876 DF,  p-value: < 2.2e-16
```

```
#Create a new variable with datetime as a factor
train$hour<-(as.POSIXlt(train$datetime))$hour
#Fit a new regression equation with the hour, weather, season, and workingday factors

better.casual_fit<-lm(formula = lcasual ~ datetime + as.factor(season) + as.factor(holiday) +
    as.factor(workingday) + as.factor(weather) + temp + atemp +
    humidity + windspeed + as.factor(hour), data = train)

better.registered_fit<-lm(formula = lregistered ~ datetime + as.factor(season) + as.factor(holiday) +
    as.factor(workingday) + as.factor(weather) + temp + atemp +
    humidity + windspeed + as.factor(hour), data = train)


summary(better.casual_fit)
```

```
##
## Call:
## lm(formula = lcasual ~ datetime + as.factor(season) + as.factor(holiday) +
##     as.factor(workingday) + as.factor(weather) + temp + atemp +
##     humidity + windspeed + as.factor(hour), data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.86310 -0.36908  0.04906  0.42591  2.39378
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -1.019e+01  5.106e-01 -19.953  < 2e-16 ***
## datetime                  8.648e-09  3.865e-10  22.376  < 2e-16 ***
## as.factor(season)2        4.719e-01  2.268e-02  20.807  < 2e-16 ***
## as.factor(season)3        1.750e-01  2.931e-02   5.972 2.42e-09 ***
## as.factor(season)4        2.525e-01  2.061e-02  12.252  < 2e-16 ***
## as.factor(holiday)1      -2.308e-01  3.784e-02  -6.099 1.10e-09 ***
## as.factor(workingday)1 -7.184e-01  1.354e-02 -53.052  < 2e-16 ***
## as.factor(weather)2      -8.831e-02  1.505e-02  -5.866 4.59e-09 ***
## as.factor(weather)3      -6.125e-01  2.544e-02 -24.075  < 2e-16 ***
## as.factor(weather)4      -3.032e-01  6.367e-01  -0.476    0.634
## temp                      3.329e-02  4.992e-03   6.668 2.71e-11 ***
## atemp                     3.466e-02  4.366e-03   7.937 2.28e-15 ***
## humidity                 -4.906e-03  4.302e-04 -11.404  < 2e-16 ***
## windspeed                -3.393e-03  8.266e-04  -4.105 4.08e-05 ***
## as.factor(hour)1         -4.129e-01  4.216e-02  -9.793  < 2e-16 ***
```

```
## as.factor(hour)2       -6.865e-01  4.232e-02 -16.222  < 2e-16 ***
## as.factor(hour)3       -1.073e+00  4.272e-02 -25.125  < 2e-16 ***
## as.factor(hour)4       -1.308e+00  4.253e-02 -30.757  < 2e-16 ***
## as.factor(hour)5       -1.165e+00  4.230e-02 -27.531  < 2e-16 ***
## as.factor(hour)6       -4.675e-01  4.225e-02 -11.065  < 2e-16 ***
## as.factor(hour)7        3.185e-01  4.220e-02   7.546 4.84e-14 ***
## as.factor(hour)8        9.360e-01  4.216e-02  22.203  < 2e-16 ***
## as.factor(hour)9        1.125e+00  4.220e-02  26.665  < 2e-16 ***
## as.factor(hour)10       1.322e+00  4.234e-02  31.234  < 2e-16 ***
## as.factor(hour)11       1.477e+00  4.258e-02  34.697  < 2e-16 ***
## as.factor(hour)12       1.539e+00  4.285e-02  35.923  < 2e-16 ***
## as.factor(hour)13       1.538e+00  4.313e-02  35.659  < 2e-16 ***
## as.factor(hour)14       1.541e+00  4.335e-02  35.539  < 2e-16 ***
## as.factor(hour)15       1.546e+00  4.341e-02  35.608  < 2e-16 ***
## as.factor(hour)16       1.568e+00  4.333e-02  36.191  < 2e-16 ***
## as.factor(hour)17       1.658e+00  4.314e-02  38.424  < 2e-16 ***
## as.factor(hour)18       1.462e+00  4.293e-02  34.059  < 2e-16 ***
## as.factor(hour)19       1.252e+00  4.254e-02  29.423  < 2e-16 ***
## as.factor(hour)20       1.016e+00  4.235e-02  23.999  < 2e-16 ***
## as.factor(hour)21       8.485e-01  4.221e-02  20.101  < 2e-16 ***
## as.factor(hour)22       6.691e-01  4.215e-02  15.874  < 2e-16 ***
## as.factor(hour)23       3.850e-01  4.213e-02   9.139  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6355 on 10849 degrees of freedom
## Multiple R-squared:  0.819,  Adjusted R-squared:  0.8184
## F-statistic:  1363 on 36 and 10849 DF,  p-value: < 2.2e-16
```

```
summary(better.registered_fit)
```

```
##
## Call:
## lm(formula = lregistered ~ datetime + as.factor(season) + as.factor(holiday) +
##     as.factor(workingday) + as.factor(weather) + temp + atemp +
##     humidity + windspeed + as.factor(hour), data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4263 -0.2998  0.0341  0.3590  2.2322
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -1.842e+01  4.772e-01 -38.608  < 2e-16 ***
## datetime                1.624e-08  3.612e-10  44.955  < 2e-16 ***
## as.factor(season)2      2.201e-01  2.120e-02  10.385  < 2e-16 ***
## as.factor(season)3      8.170e-02  2.739e-02   2.983 0.002862 **
## as.factor(season)4      1.773e-01  1.926e-02   9.206  < 2e-16 ***
## as.factor(holiday)1    -3.760e-02  3.536e-02  -1.063 0.287683
## as.factor(workingday)1  8.153e-02  1.266e-02   6.443 1.22e-10 ***
## as.factor(weather)2    -4.967e-02  1.407e-02  -3.531 0.000416 ***
## as.factor(weather)3    -5.009e-01  2.378e-02 -21.069  < 2e-16 ***
## as.factor(weather)4    -2.654e-01  5.950e-01  -0.446 0.655611
## temp                    1.339e-02  4.666e-03   2.871 0.004101 **
## atemp                   1.103e-02  4.081e-03   2.703 0.006891 **
```

```
## humidity               -2.198e-03  4.021e-04  -5.466 4.72e-08 ***
## windspeed              -2.997e-03  7.725e-04  -3.880 0.000105 ***
## as.factor(hour)1       -6.193e-01  3.941e-02 -15.715  < 2e-16 ***
## as.factor(hour)2       -1.135e+00  3.955e-02 -28.698  < 2e-16 ***
## as.factor(hour)3       -1.585e+00  3.992e-02 -39.702  < 2e-16 ***
## as.factor(hour)4       -1.839e+00  3.975e-02 -46.265  < 2e-16 ***
## as.factor(hour)5       -8.045e-01  3.954e-02 -20.349  < 2e-16 ***
## as.factor(hour)6        3.710e-01  3.948e-02   9.395  < 2e-16 ***
## as.factor(hour)7        1.341e+00  3.944e-02  34.006  < 2e-16 ***
## as.factor(hour)8        1.968e+00  3.940e-02  49.958  < 2e-16 ***
## as.factor(hour)9        1.585e+00  3.944e-02  40.188  < 2e-16 ***
## as.factor(hour)10       1.158e+00  3.957e-02  29.259  < 2e-16 ***
## as.factor(hour)11       1.259e+00  3.979e-02  31.628  < 2e-16 ***
## as.factor(hour)12       1.475e+00  4.004e-02  36.833  < 2e-16 ***
## as.factor(hour)13       1.433e+00  4.031e-02  35.542  < 2e-16 ***
## as.factor(hour)14       1.319e+00  4.052e-02  32.544  < 2e-16 ***
## as.factor(hour)15       1.392e+00  4.057e-02  34.312  < 2e-16 ***
## as.factor(hour)16       1.712e+00  4.049e-02  42.271  < 2e-16 ***
## as.factor(hour)17       2.173e+00  4.032e-02  53.908  < 2e-16 ***
## as.factor(hour)18       2.113e+00  4.012e-02  52.678  < 2e-16 ***
## as.factor(hour)19       1.817e+00  3.976e-02  45.703  < 2e-16 ***
## as.factor(hour)20       1.511e+00  3.958e-02  38.179  < 2e-16 ***
## as.factor(hour)21       1.254e+00  3.945e-02  31.787  < 2e-16 ***
## as.factor(hour)22       1.009e+00  3.939e-02  25.615  < 2e-16 ***
## as.factor(hour)23       6.084e-01  3.937e-02  15.453  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5939 on 10849 degrees of freedom
## Multiple R-squared:  0.8205, Adjusted R-squared:  0.8199
## F-statistic:  1377 on 36 and 10849 DF,  p-value: < 2.2e-16
```

Here we see drastic improvements in our multiple R-squared for both lregistered and lcasual.

With all of the hour factors, it would most likely be useful to create factors that go along with the different times of day since we know that there will be fewer users late at night and in the middle of the night and we will see an increase in users during the day.

As far as variables that could be useful it is important to note which variables do not do much explaining on their own. One such case is the variable season which does give a broad idea as to what the weather and temperature are but does not give specifics. There are certainly warmer days in the beginning of fall compared to the end of fall and there will be days in the spring that do not make for ideal biking conditions, so interacting the atemp and season as well as the season and time of day should result in something more accurate. Also important to include are the interactions between the atemp and time of day as well as the atemp and humidity.

```
#Create a variable to represent different times in the day
breaks<-c(0,5,12,21,23)
labels<-c("middle of night","morning","afternoon","night")
train$tod<-cut(train$hour,breaks,labels,include.lowest = TRUE)

#Fit new regression equation with interactions
interaction.casual_fit<-lm(formula = lcasual ~ datetime + as.factor(season) + as.factor(holiday) +
    as.factor(workingday) + as.factor(weather) + temp + atemp +
    humidity + windspeed + as.factor(hour) + as.factor(season):atemp +
    as.factor(tod):atemp + humidity:as.factor(season) + as.factor(tod):humidity +
```

```
    as.factor(tod):windspeed, data = train)

summary(interaction.casual_fit)
```

```
##
## Call:
## lm(formula = lcasual ~ datetime + as.factor(season) + as.factor(holiday) +
##     as.factor(workingday) + as.factor(weather) + temp + atemp +
##     humidity + windspeed + as.factor(hour) + as.factor(season):atemp +
##     as.factor(tod):atemp + humidity:as.factor(season) + as.factor(tod):humidity +
##     as.factor(tod):windspeed, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8094 -0.3493  0.0480  0.3950  2.2914
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -9.833e+00  4.904e-01 -20.051  < 2e-16
## datetime                         8.264e-09  3.707e-10  22.294  < 2e-16
## as.factor(season)2               1.757e+00  9.249e-02  19.002  < 2e-16
## as.factor(season)3               2.648e+00  1.486e-01  17.818  < 2e-16
## as.factor(season)4               7.193e-01  7.958e-02   9.039  < 2e-16
## as.factor(holiday)1             -1.778e-01  3.604e-02  -4.934 8.17e-07
## as.factor(workingday)1          -7.233e-01  1.287e-02 -56.201  < 2e-16
## as.factor(weather)2             -8.624e-02  1.432e-02  -6.023 1.77e-09
## as.factor(weather)3             -6.187e-01  2.428e-02 -25.482  < 2e-16
## as.factor(weather)4             -5.829e-02  6.037e-01  -0.097  0.92309
## temp                             1.374e-02  5.242e-03   2.621  0.00877
## atemp                            4.414e-02  4.949e-03   8.919  < 2e-16
## humidity                         2.113e-03  9.472e-04   2.231  0.02572
## windspeed                       -5.174e-04  1.720e-03  -0.301  0.76359
## as.factor(hour)1                -4.309e-01  3.997e-02 -10.782  < 2e-16
## as.factor(hour)2                -7.180e-01  4.015e-02 -17.881  < 2e-16
## as.factor(hour)3                -1.116e+00  4.060e-02 -27.479  < 2e-16
## as.factor(hour)4                -1.371e+00  4.057e-02 -33.797  < 2e-16
## as.factor(hour)5                -1.240e+00  4.037e-02 -30.707  < 2e-16
## as.factor(hour)6                -7.405e-01  9.862e-02  -7.508 6.45e-14
## as.factor(hour)7                 5.557e-02  9.828e-02   0.565  0.57183
## as.factor(hour)8                 6.800e-01  9.796e-02   6.942 4.10e-12
## as.factor(hour)9                 8.744e-01  9.744e-02   8.973  < 2e-16
## as.factor(hour)10                1.074e+00  9.672e-02  11.103  < 2e-16
## as.factor(hour)11                1.225e+00  9.603e-02  12.757  < 2e-16
## as.factor(hour)12                1.279e+00  9.567e-02  13.374  < 2e-16
## as.factor(hour)13                7.721e-01  9.195e-02   8.397  < 2e-16
## as.factor(hour)14                7.616e-01  9.211e-02   8.268  < 2e-16
## as.factor(hour)15                7.631e-01  9.219e-02   8.277  < 2e-16
## as.factor(hour)16                7.885e-01  9.223e-02   8.548  < 2e-16
## as.factor(hour)17                8.937e-01  9.219e-02   9.694  < 2e-16
## as.factor(hour)18                7.098e-01  9.214e-02   7.703 1.45e-14
## as.factor(hour)19                5.113e-01  9.216e-02   5.549 2.95e-08
## as.factor(hour)20                2.846e-01  9.213e-02   3.089  0.00201
## as.factor(hour)21                1.265e-01  9.224e-02   1.371  0.17040
## as.factor(hour)22               -2.581e-01  1.297e-01  -1.991  0.04656
```

```
## as.factor(hour)23                 -5.427e-01  1.298e-01  -4.181 2.93e-05
## as.factor(season)2:atemp          -4.686e-02  2.975e-03 -15.753  < 2e-16
## as.factor(season)3:atemp          -7.588e-02  3.672e-03 -20.663  < 2e-16
## as.factor(season)4:atemp          -6.527e-03  2.965e-03  -2.201  0.02773
## atemp:as.factor(tod)morning        2.877e-02  2.029e-03  14.180  < 2e-16
## atemp:as.factor(tod)afternoon      4.761e-02  2.033e-03  23.421  < 2e-16
## atemp:as.factor(tod)night          5.389e-02  2.997e-03  17.983  < 2e-16
## as.factor(season)2:humidity       -5.310e-03  8.502e-04  -6.246 4.37e-10
## as.factor(season)3:humidity       -6.449e-03  1.031e-03  -6.254 4.16e-10
## as.factor(season)4:humidity       -7.240e-03  8.757e-04  -8.267  < 2e-16
## humidity:as.factor(tod)morning    -5.140e-03  1.042e-03  -4.934 8.18e-07
## humidity:as.factor(tod)afternoon  -4.980e-03  9.621e-04  -5.176 2.31e-07
## humidity:as.factor(tod)night      -2.794e-03  1.441e-03  -1.939  0.05251
## windspeed:as.factor(tod)morning   -2.457e-03  2.189e-03  -1.122  0.26168
## windspeed:as.factor(tod)afternoon -2.442e-03  2.060e-03  -1.185  0.23597
## windspeed:as.factor(tod)night     -8.639e-03  3.243e-03  -2.664  0.00774
##
## (Intercept)                       ***
## datetime                          ***
## as.factor(season)2                ***
## as.factor(season)3                ***
## as.factor(season)4                ***
## as.factor(holiday)1               ***
## as.factor(workingday)1            ***
## as.factor(weather)2               ***
## as.factor(weather)3               ***
## as.factor(weather)4
## temp                              **
## atemp                             ***
## humidity                          *
## windspeed
## as.factor(hour)1                  ***
## as.factor(hour)2                  ***
## as.factor(hour)3                  ***
## as.factor(hour)4                  ***
## as.factor(hour)5                  ***
## as.factor(hour)6                  ***
## as.factor(hour)7
## as.factor(hour)8                  ***
## as.factor(hour)9                  ***
## as.factor(hour)10                 ***
## as.factor(hour)11                 ***
## as.factor(hour)12                 ***
## as.factor(hour)13                 ***
## as.factor(hour)14                 ***
## as.factor(hour)15                 ***
## as.factor(hour)16                 ***
## as.factor(hour)17                 ***
## as.factor(hour)18                 ***
## as.factor(hour)19                 ***
## as.factor(hour)20                 **
## as.factor(hour)21
## as.factor(hour)22                 *
## as.factor(hour)23                 ***
```

```
## as.factor(season)2:atemp          ***
## as.factor(season)3:atemp          ***
## as.factor(season)4:atemp          *
## atemp:as.factor(tod)morning       ***
## atemp:as.factor(tod)afternoon     ***
## atemp:as.factor(tod)night         ***
## as.factor(season)2:humidity       ***
## as.factor(season)3:humidity       ***
## as.factor(season)4:humidity       ***
## humidity:as.factor(tod)morning    ***
## humidity:as.factor(tod)afternoon  ***
## humidity:as.factor(tod)night      .
## windspeed:as.factor(tod)morning
## windspeed:as.factor(tod)afternoon
## windspeed:as.factor(tod)night     **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.602 on 10834 degrees of freedom
## Multiple R-squared:  0.8378, Adjusted R-squared:  0.837
## F-statistic:  1097 on 51 and 10834 DF,  p-value: < 2.2e-16
```

```
interaction.registered_fit<-lm(formula = lregistered ~ datetime + as.factor(season) + as.factor(holiday)
    as.factor(workingday) + as.factor(weather) + temp + atemp +
    humidity + windspeed + as.factor(hour) + as.factor(season):atemp +
    as.factor(tod):atemp + as.factor(season):humidity, data = train)


summary(interaction.registered_fit)
```

```
##
## Call:
## lm(formula = lregistered ~ datetime + as.factor(season) + as.factor(holiday) +
##     as.factor(workingday) + as.factor(weather) + temp + atemp +
##     humidity + windspeed + as.factor(hour) + as.factor(season):atemp +
##     as.factor(tod):atemp + as.factor(season):humidity, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5098 -0.2923  0.0266  0.3490  2.2593
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -1.827e+01  4.769e-01 -38.306  < 2e-16 ***
## datetime                      1.598e-08  3.630e-10  44.025  < 2e-16 ***
## as.factor(season)2            3.273e-01  9.058e-02   3.614 0.000303 ***
## as.factor(season)3            1.263e+00  1.455e-01   8.676  < 2e-16 ***
## as.factor(season)4            5.456e-01  7.791e-02   7.003 2.65e-12 ***
## as.factor(holiday)1          -1.111e-02  3.530e-02  -0.315 0.753041
## as.factor(workingday)1        8.220e-02  1.260e-02   6.523 7.19e-11 ***
## as.factor(weather)2          -4.987e-02  1.398e-02  -3.566 0.000364 ***
## as.factor(weather)3          -4.932e-01  2.367e-02 -20.840  < 2e-16 ***
## as.factor(weather)4          -2.260e-01  5.913e-01  -0.382 0.702292
## temp                          7.037e-04  5.111e-03   0.138 0.890484
## atemp                         3.085e-02  4.826e-03   6.392 1.71e-10 ***
```

```
## humidity                       -1.458e-03  6.258e-04  -2.330 0.019851 *
## windspeed                      -2.676e-03  7.824e-04  -3.419 0.000630 ***
## as.factor(hour)1               -6.190e-01  3.914e-02 -15.816  < 2e-16 ***
## as.factor(hour)2               -1.136e+00  3.929e-02 -28.916  < 2e-16 ***
## as.factor(hour)3               -1.587e+00  3.966e-02 -40.003  < 2e-16 ***
## as.factor(hour)4               -1.841e+00  3.952e-02 -46.584  < 2e-16 ***
## as.factor(hour)5               -8.064e-01  3.934e-02 -20.498  < 2e-16 ***
## as.factor(hour)6                4.376e-01  5.789e-02   7.559 4.40e-14 ***
## as.factor(hour)7                1.413e+00  5.792e-02  24.398  < 2e-16 ***
## as.factor(hour)8                2.046e+00  5.819e-02  35.161  < 2e-16 ***
## as.factor(hour)9                1.668e+00  5.863e-02  28.451  < 2e-16 ***
## as.factor(hour)10               1.246e+00  5.914e-02  21.061  < 2e-16 ***
## as.factor(hour)11               1.349e+00  5.980e-02  22.553  < 2e-16 ***
## as.factor(hour)12               1.564e+00  6.036e-02  25.916  < 2e-16 ***
## as.factor(hour)13               1.358e+00  6.065e-02  22.388  < 2e-16 ***
## as.factor(hour)14               1.240e+00  6.101e-02  20.332  < 2e-16 ***
## as.factor(hour)15               1.312e+00  6.113e-02  21.464  < 2e-16 ***
## as.factor(hour)16               1.632e+00  6.099e-02  26.758  < 2e-16 ***
## as.factor(hour)17               2.097e+00  6.060e-02  34.600  < 2e-16 ***
## as.factor(hour)18               2.039e+00  6.028e-02  33.824  < 2e-16 ***
## as.factor(hour)19               1.744e+00  5.980e-02  29.171  < 2e-16 ***
## as.factor(hour)20               1.439e+00  5.951e-02  24.180  < 2e-16 ***
## as.factor(hour)21               1.183e+00  5.928e-02  19.963  < 2e-16 ***
## as.factor(hour)22               7.079e-01  7.689e-02   9.207  < 2e-16 ***
## as.factor(hour)23               3.119e-01  7.607e-02   4.101 4.15e-05 ***
## as.factor(season)2:atemp       -7.743e-03  2.912e-03  -2.659 0.007842 **
## as.factor(season)3:atemp       -3.514e-02  3.592e-03  -9.782  < 2e-16 ***
## as.factor(season)4:atemp       -8.426e-03  2.897e-03  -2.909 0.003637 **
## atemp:as.factor(tod)morning    -3.303e-03  1.945e-03  -1.698 0.089515 .
## atemp:as.factor(tod)afternoon   3.019e-03  1.954e-03   1.545 0.122406
## atemp:as.factor(tod)night       1.300e-02  2.876e-03   4.521 6.23e-06 ***
## as.factor(season)2:humidity    -1.890e-04  8.326e-04  -0.227 0.820440
## as.factor(season)3:humidity    -3.034e-03  1.010e-03  -3.005 0.002665 **
## as.factor(season)4:humidity    -3.750e-03  8.561e-04  -4.380 1.20e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5898 on 10840 degrees of freedom
## Multiple R-squared:  0.8231, Adjusted R-squared:  0.8224
## F-statistic:  1121 on 45 and 10840 DF,  p-value: < 2.2e-16
```

Question 5

First to clean up our equation for lcasual by stepwise regression

```
step(interaction.casual_fit,direction = "both")
```

```
## Start:  AIC=-10997.01
## lcasual ~ datetime + as.factor(season) + as.factor(holiday) +
##     as.factor(workingday) + as.factor(weather) + temp + atemp +
##     humidity + windspeed + as.factor(hour) + as.factor(season):atemp +
##     as.factor(tod):atemp + humidity:as.factor(season) + as.factor(tod):humidity +
##     as.factor(tod):windspeed
##
##                              Df Sum of Sq    RSS      AIC
## <none>                                    3926.4 -10997.0
```

```
## - windspeed:as.factor(tod)    3      2.57 3929.0 -10995.9
## - temp                        1      2.49 3928.9 -10992.1
## - as.factor(holiday)          1      8.82 3935.2 -10974.6
## - humidity:as.factor(tod)     3     11.66 3938.1 -10970.7
## - as.factor(season):humidity  3     30.09 3956.5 -10919.9
## - datetime                    1    180.12 4106.5 -10510.7
## - as.factor(season):atemp     3    221.86 4148.3 -10404.6
## - atemp:as.factor(tod)        3    230.78 4157.2 -10381.3
## - as.factor(weather)          3    236.14 4162.6 -10367.2
## - as.factor(workingday)       1   1144.73 5071.1  -8213.9
## - as.factor(hour)            23   2144.40 6070.8  -6299.3
```

```
##
## Call:
## lm(formula = lcasual ~ datetime + as.factor(season) + as.factor(holiday) +
##     as.factor(workingday) + as.factor(weather) + temp + atemp +
##     humidity + windspeed + as.factor(hour) + as.factor(season):atemp +
##     as.factor(tod):atemp + humidity:as.factor(season) + as.factor(tod):humidity +
##     as.factor(tod):windspeed, data = train)
##
## Coefficients:
##                       (Intercept)                          datetime
##                         -9.833e+00                         8.264e-09
##                 as.factor(season)2                 as.factor(season)3
##                          1.757e+00                          2.648e+00
##                 as.factor(season)4                as.factor(holiday)1
##                          7.193e-01                         -1.778e-01
##             as.factor(workingday)1                as.factor(weather)2
##                         -7.233e-01                         -8.624e-02
##                 as.factor(weather)3                as.factor(weather)4
##                         -6.187e-01                         -5.829e-02
##                                temp                              atemp
##                          1.374e-02                          4.414e-02
##                            humidity                          windspeed
##                          2.113e-03                         -5.174e-04
##                   as.factor(hour)1                   as.factor(hour)2
##                         -4.309e-01                         -7.180e-01
##                   as.factor(hour)3                   as.factor(hour)4
##                         -1.116e+00                         -1.371e+00
##                   as.factor(hour)5                   as.factor(hour)6
##                         -1.240e+00                         -7.405e-01
##                   as.factor(hour)7                   as.factor(hour)8
##                          5.557e-02                          6.800e-01
##                   as.factor(hour)9                  as.factor(hour)10
##                          8.744e-01                          1.074e+00
##                  as.factor(hour)11                  as.factor(hour)12
##                          1.225e+00                          1.279e+00
##                  as.factor(hour)13                  as.factor(hour)14
##                          7.721e-01                          7.616e-01
##                  as.factor(hour)15                  as.factor(hour)16
##                          7.631e-01                          7.885e-01
##                  as.factor(hour)17                  as.factor(hour)18
##                          8.937e-01                          7.098e-01
##                  as.factor(hour)19                  as.factor(hour)20
```

```
##                          5.113e-01                          2.846e-01
##                 as.factor(hour)21                 as.factor(hour)22
##                          1.265e-01                         -2.581e-01
##                 as.factor(hour)23          as.factor(season)2:atemp
##                         -5.427e-01                         -4.686e-02
##          as.factor(season)3:atemp          as.factor(season)4:atemp
##                         -7.588e-02                         -6.527e-03
##       atemp:as.factor(tod)morning     atemp:as.factor(tod)afternoon
##                          2.877e-02                          4.761e-02
##         atemp:as.factor(tod)night        as.factor(season)2:humidity
##                          5.389e-02                         -5.310e-03
##       as.factor(season)3:humidity        as.factor(season)4:humidity
##                         -6.449e-03                         -7.240e-03
##    humidity:as.factor(tod)morning   humidity:as.factor(tod)afternoon
##                         -5.140e-03                         -4.980e-03
##      humidity:as.factor(tod)night    windspeed:as.factor(tod)morning
##                         -2.794e-03                         -2.457e-03
## windspeed:as.factor(tod)afternoon      windspeed:as.factor(tod)night
##                         -2.442e-03                         -8.639e-03
```

There is nothing to change in our interaction.casual_fit equation. The current equation already has the lowest AIC.

For the registered_fit equation

```
step(interaction.registered_fit,direction="both")
```

```
## Start:  AIC=-11449.34
## lregistered ~ datetime + as.factor(season) + as.factor(holiday) +
##     as.factor(workingday) + as.factor(weather) + temp + atemp +
##     humidity + windspeed + as.factor(hour) + as.factor(season):atemp +
##     as.factor(tod):atemp + as.factor(season):humidity
##
##                              Df Sum of Sq    RSS      AIC
## - temp                        1       0.0 3770.8 -11451.3
## - as.factor(holiday)          1       0.0 3770.8 -11451.2
## <none>                                    3770.8 -11449.3
## - windspeed                   1       4.1 3774.8 -11439.6
## - as.factor(season):humidity  3       9.7 3780.5 -11427.3
## - atemp:as.factor(tod)        3      13.8 3784.5 -11415.7
## - as.factor(workingday)       1      14.8 3785.6 -11408.7
## - as.factor(season):atemp     3      33.3 3804.1 -11359.5
## - as.factor(weather)          3     153.3 3924.1 -11021.5
## - datetime                    1     674.2 4445.0  -9660.6
## - as.factor(hour)            23    3399.0 7169.8  -4500.0
##
## Step:  AIC=-11451.32
## lregistered ~ datetime + as.factor(season) + as.factor(holiday) +
##     as.factor(workingday) + as.factor(weather) + atemp + humidity +
##     windspeed + as.factor(hour) + as.factor(season):atemp + atemp:as.factor(tod) +
##     as.factor(season):humidity
##
##                              Df Sum of Sq    RSS      AIC
## - as.factor(holiday)          1       0.0 3770.8 -11453.2
## <none>                                    3770.8 -11451.3
## + temp                        1       0.0 3770.8 -11449.3
```

```
## - windspeed                    1       4.2 3775.0 -11441.1
## - as.factor(season):humidity  3       9.9 3780.7 -11428.8
## - atemp:as.factor(tod)        3      13.8 3784.5 -11417.6
## - as.factor(workingday)       1      14.8 3785.6 -11410.6
## - as.factor(season):atemp     3      37.1 3807.9 -11350.6
## - as.factor(weather)          3     153.5 3924.3 -11023.0
## - datetime                    1     676.5 4447.3  -9656.9
## - as.factor(hour)            23    3399.7 7170.4  -4501.0
##
## Step:  AIC=-11453.23
## lregistered ~ datetime + as.factor(season) + as.factor(workingday) +
##     as.factor(weather) + atemp + humidity + windspeed + as.factor(hour) +
##     as.factor(season):atemp + atemp:as.factor(tod) + as.factor(season):humidity
##
##                              Df Sum of Sq    RSS      AIC
## <none>                                     3770.8 -11453.2
## + as.factor(holiday)          1       0.0 3770.8 -11451.3
## + temp                        1       0.0 3770.8 -11451.2
## - windspeed                   1       4.2 3775.0 -11443.0
## - as.factor(season):humidity  3       9.9 3780.7 -11430.6
## - atemp:as.factor(tod)        3      13.8 3784.6 -11419.5
## - as.factor(workingday)       1      16.2 3787.0 -11408.6
## - as.factor(season):atemp     3      37.4 3808.2 -11351.8
## - as.factor(weather)          3     153.5 3924.3 -11025.0
## - datetime                    1     676.5 4447.3  -9658.9
## - as.factor(hour)            23    3399.7 7170.5  -4503.0
##
## Call:
## lm(formula = lregistered ~ datetime + as.factor(season) + as.factor(workingday) +
##     as.factor(weather) + atemp + humidity + windspeed + as.factor(hour) +
##     as.factor(season):atemp + atemp:as.factor(tod) + as.factor(season):humidity,
##     data = train)
##
## Coefficients:
##                   (Intercept)                      datetime
##                     -1.827e+01                     1.598e-08
##            as.factor(season)2             as.factor(season)3
##                      3.267e-01                     1.273e+00
##            as.factor(season)4         as.factor(workingday)1
##                      5.462e-01                     8.323e-02
##           as.factor(weather)2            as.factor(weather)3
##                     -4.993e-02                    -4.930e-01
##           as.factor(weather)4                         atemp
##                     -2.266e-01                     3.147e-02
##                       humidity                     windspeed
##                     -1.456e-03                    -2.652e-03
##              as.factor(hour)1               as.factor(hour)2
##                     -6.190e-01                    -1.136e+00
##              as.factor(hour)3               as.factor(hour)4
##                     -1.587e+00                    -1.841e+00
##              as.factor(hour)5               as.factor(hour)6
##                     -8.064e-01                     4.376e-01
##              as.factor(hour)7               as.factor(hour)8
```

```
##                     1.413e+00                     2.046e+00
##              as.factor(hour)9             as.factor(hour)10
##                     1.668e+00                     1.246e+00
##             as.factor(hour)11             as.factor(hour)12
##                     1.349e+00                     1.565e+00
##             as.factor(hour)13             as.factor(hour)14
##                     1.358e+00                     1.241e+00
##             as.factor(hour)15             as.factor(hour)16
##                     1.312e+00                     1.632e+00
##             as.factor(hour)17             as.factor(hour)18
##                     2.097e+00                     2.039e+00
##             as.factor(hour)19             as.factor(hour)20
##                     1.745e+00                     1.439e+00
##             as.factor(hour)21             as.factor(hour)22
##                     1.183e+00                     7.080e-01
##             as.factor(hour)23      as.factor(season)2:atemp
##                     3.120e-01                    -7.713e-03
##      as.factor(season)3:atemp      as.factor(season)4:atemp
##                    -3.535e-02                    -8.458e-03
##   atemp:as.factor(tod)morning atemp:as.factor(tod)afternoon
##                    -3.305e-03                     3.020e-03
##     atemp:as.factor(tod)night    as.factor(season)2:humidity
##                     1.300e-02                    -1.879e-04
##   as.factor(season)3:humidity    as.factor(season)4:humidity
##                    -3.080e-03                    -3.752e-03
```

```
#Fit new registered equation based on the stepwise regression
interaction.registered_fit<-lm(formula = lregistered ~ datetime + as.factor(season) + as.factor(working
    as.factor(weather) + atemp + humidity + windspeed + as.factor(hour) +
    as.factor(season):atemp + atemp:as.factor(tod) + as.factor(season):humidity,
    data = train)
```

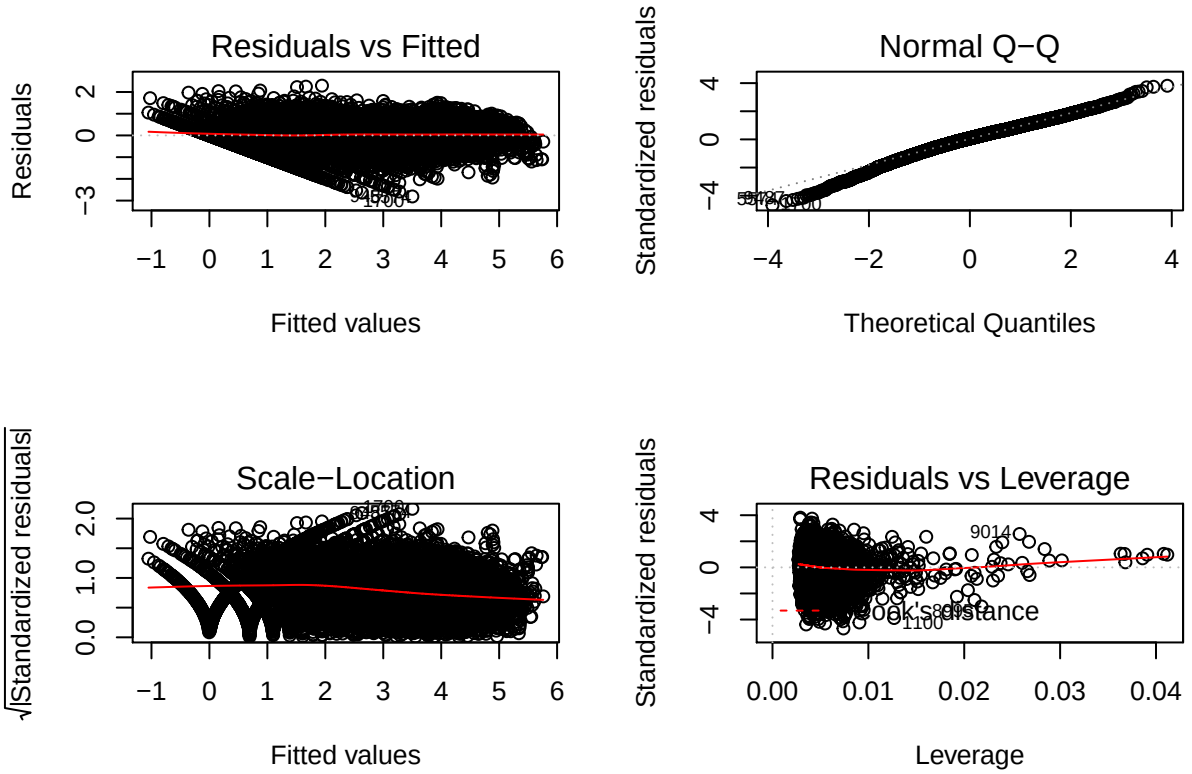
The temp and as.factor(holiday) variables have been dropped from the interacted.registered_fit equation to attain an equation with the lowest AIC.

Question 6

```
par(mfrow=c(2,2))
plot(interaction.casual_fit)
```

```
## Warning: not plotting observations with leverage one:
##   5632
```

```
## Warning: not plotting observations with leverage one:
##   5632
```



The first plot resembles a random scatter but there is an odd slice of missing values in the bottom left hand side. However there is nothing that is alarmingly out of the norm here.

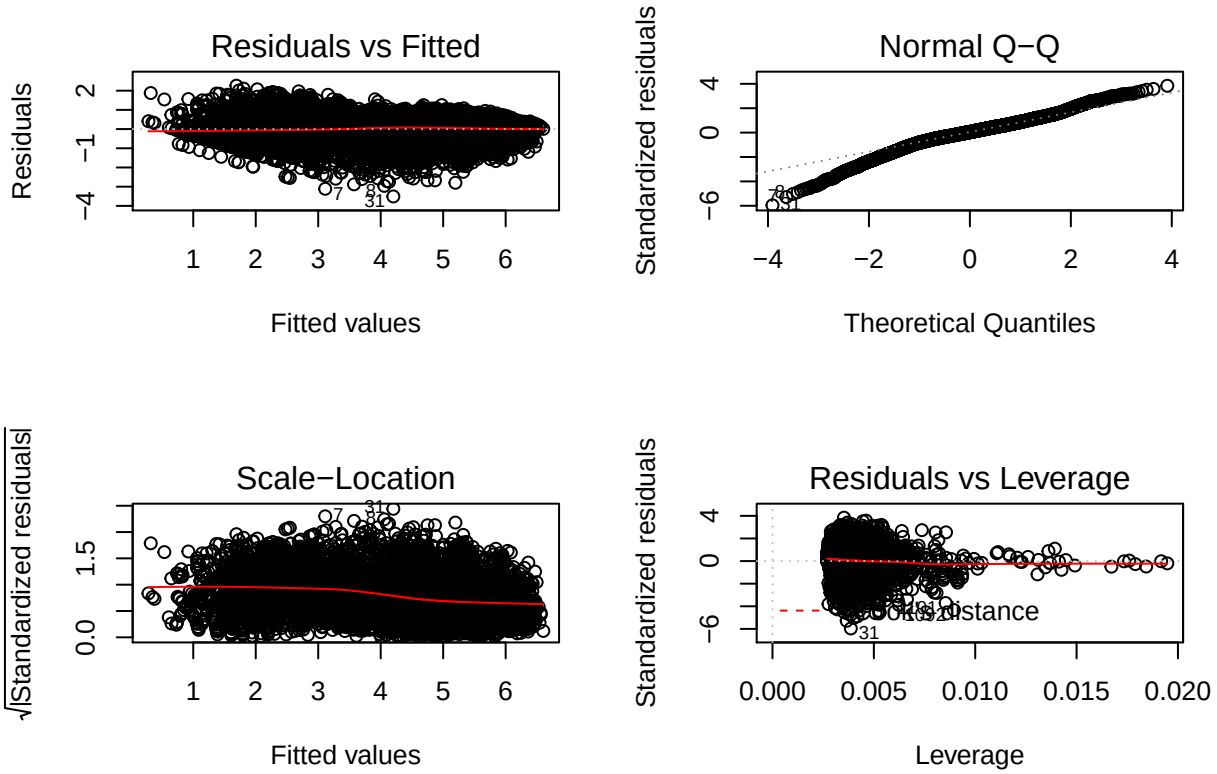
The normal assumption holds on the second plot as the points fall along the plotted line.

The third plot does not necessarily indicate any increasing or decreasing so we cannot assume heterscedasticity.

```
par(mfrow=c(2,2))
plot(interaction.registered_fit)
```

```
## Warning: not plotting observations with leverage one:
##   5632

## Warning: not plotting observations with leverage one:
##   5632
```



The first plot resembles a random scatter but the bottom lower left portion of the plot is missing values.

The Normal Q-Q plot shows the points along the plotted line.

The third plot does not show any signs of increases or decreases, so we can not assume any heterscadicity.

Question 7

Prediction for registered and

```
#First read in test data
test <- read_csv("~/stat28/projects/data/test.csv")
```

```
## Parsed with column specification:
## cols(
##   datetime = col_datetime(format = ""),
##   season = col_integer(),
##   holiday = col_integer(),
##   workingday = col_integer(),
##   weather = col_integer(),
##   temp = col_double(),
##   atemp = col_double(),
##   humidity = col_integer(),
##   windspeed = col_double()
## )
```

```
#Add the previous variables from the train dataset
#hour
test$hour<-(as.POSIXlt(test$datetime))$hour
#Time of day labels
breaks<-c(0,5,12,21,23)
labels<-c("middle of night","morning","afternoon","night")
```

```
test$tod<-cut(test$hour,breaks,labels,include.lowest = TRUE)
```

```
#Predictions
predicted_registered<-exp(predict(interaction.registered_fit,test))
predicted_casual<-exp(predict(interaction.casual_fit,test))
predicted_count=predicted_casual+predicted_registered
#Prediction data frame
prediction<-data.frame("datetime"=test$datetime,
                       "count"=predicted_count)
```

```
#Write the prediction data frame into a csv
write.csv(prediction,file="Subm.csv", row.names = FALSE)
```

Kaggle score: 0.62149

Question 8

Certainly far from the top but a score that was better than expected. In the future I would look more into interacting more variables and possibly start with all combinations and eventually do stepwise regression to see if there were any interactions that would give me a better regression equation and more accurate predictions.