Group Name: DDS: Dust of Data Science.
Group Members: Seojin Han, Seungwoo Lee, Insu Kim, Ju Ho Yoon

Group Project Part 3: Preliminary Analysis

In coming up with our analysis, first we have collected the various data sets from the KOSIS, the Korean Statistical Information Service. From each of the data sets we have collected, we organized all the data into an excel and have fined them into the time line of 2005 to 2013 for all of the data we collected since other years were not all present. After organizing them into a one collective data, we have set the level of fine dusts in Seoul as our dependent variable, and set the potential factors as our independent variables (x1, x2, x3…etc). In doing so, we realize that the levels of different factors might show certain trends due to other lurking variables or other factors. Hence, to control for such issue, we have generally controlled for the potential factors by dividing each of the factors to the corresponding population of Seoul for every year as by the following in excel:

Figure 1:

Finedust(㎍/㎥)	car registration	bikeroad(len,km)	bikeroad(num)	subway users	GRDP(bil,won)	street tree	green belt(㎡)	coal power generation(mil.kwh)	nuclear energy generation	Construction Expenditure	Population
76	2691431	554	271	2230783000				118,022	119,103		10041
69	2776536	587	281	2249226000	194891			120,277	129,672		10029
61	2779841	616	282	2300735000	198925	276779	8185144	127,164	130,714		10036
58	2808771	629	297	2277298000	208899	279461	8708112	133,658	146,779	11240735	10011
60	2856857	648	313	2269410000	231223	280243	7738173	139,205	148,749	14840812	10035
61	2933286	715	358	2267676000	249484	280499	9939148	154,674	142,937	17405539	10042
55	2949211	728	358	2293848000	262999	279442	13085443	173,508	150,958	21242916	10081
56	2954704	764	390	2293042000	273198	279672	12888759	193,216	147,771	18982632	10103
52	2981400	844	422	2446519000	289718	283609	13418116	193,769	148,596	17587044	10050
47	2977599	804	399	2518165000	303812	284305	13772994	200,124	154,723	14480078	10041
43	2969184	666	421	2559655000	313478	284476	13926414	180,752	150,327	9980630	9983
47	2973877	707	365	2619529000	318607	284498	14148342	200,444	138,784	10788117	9926
46	3013541			2660907000		293389	14387077			11987392	9890
	Index(/population)										
76	0.2680307189	0.00005517102919	0.00002698799443	222.1563069	0	0	0	0.01175342095	0.0118610741	0	
69	0.2768290094	0.00005852566959	0.00002801654711	224.2546128	0.01943122022	0	0	0.01199197949	0.01292868931	0	
61	0.2769802957	0.00006137756158	0.00002809816942	229.2427015	0.01982066792	0.02757795473	0.8155587336	0.01267048091	0.013024199	0	
58	0.2805593945	0.00006282885261	0.00002966640576	227.4722105	0.02086627103	0.02791448963	0.8698262088	0.01335068169	0.01466129755	1.122802039	
60	0.2846785925	0.00006457156517	0.00003118966034	226.1409811	0.0230407886	0.02792550793	0.771089417	0.01387142705	0.01482246257	1.478849474	
61	0.2920989801	0.00007120027532	0.00003564992806	225.8170008	0.02484381747	0.02793231612	0.9897483553	0.01540256138	0.01423378147	1.733257579	
55	0.29255094	0.00007221493625	0.00003551229008	227.5413284	0.02608853849	0.02771962392	1.298028066	0.01721135873	0.01497448125	2.107219539	
56	0.2924328416	0.00007561457627	0.0000385990638	226.9468576	0.02703894111	0.02767968557	1.275625721	0.01912296593	0.01462518528	1.878748266	
52	0.296641722	0.00008397585475	0.00004198792738	243.422422	0.02882620461	0.02821837463	1.335068436	0.01927952299	0.0147849243	1.749866176	
47	0.2965209465	0.00008006546248	0.00003973397951	250.7687131	0.03025478643	0.02831220312	1.371568575	0.01992913012	0.01540792108	1.441982763	
43	0.2974106444	0.0000667104124	0.00004216979523	256.3898508	0.03139976976	0.02849476168	1.394950182	0.01810516586	0.01505762187	0.9997176327	
47	0.2995932154	0.00007122433217	0.00003677069482	263.8956204	0.0320969884	0.02866079215	1.42532703	0.02019305521	0.01398132633	1.086812487	
46	0.3046855008	0	0	269.032272	0	0.02966323484	1.454612285	0	0	1.211990988	

On the bottom half of the excel sheet are the data that we will be using to build upon our hypothesis. Our null hypothesis is that we expect to see no correlation in any of

the factors to the fine dust level in Seoul from 2005 to 2013, whereas our alternate hypothesis will be that we expect to see a correlation in any of the factors to the fine dust level in Seoul from 2005 to 2013. We would like to build upon testing our hypothesis by first using 'R' to compare and contrast the different correlation values in each of the factors. Hence, we have exported the above file into R as the following figure:

Figure 2:

```
> data <- read.csv(file.choose(),header=T)
> data
  finedust car_regist bikeroad_lg bikeroad_nu sub_users       GRDP
1       58  0.2805594    6.28e-05  2.9700e-05  227.4722 0.02086627
2       60  0.2846786    6.46e-05  3.1200e-05  226.1410 0.02304079
3       61  0.2920990    7.12e-05  3.5600e-05  225.8170 0.02484382
4       55  0.2925509    7.22e-05  3.5500e-05  227.5413 0.02608854
5       56  0.2924328    7.56e-05  3.8600e-05  226.9469 0.02703894
6       52  0.2966417    8.40e-05  4.2000e-05  243.4224 0.02882621
7       47  0.2965209    8.01e-05  3.9734e-05  250.7687 0.03025479
8       43  0.2974106    6.67e-05  4.2200e-05  256.3899 0.03139977
9       47  0.2995932    7.12e-05  3.6800e-05  263.8956 0.03209699
        tree greenbelt  coalpower nuclearenergy construction
1 0.02791449 0.8698262 0.01335068    0.01466130    1.1228020
2 0.02792551 0.7710894 0.01387143    0.01482246    1.4788495
3 0.02793232 0.9897484 0.01540256    0.01423378    1.7332576
4 0.02771962 1.2980281 0.01721136    0.01497448    2.1072195
5 0.02767969 1.2756257 0.01912297    0.01462519    1.8787483
6 0.02821838 1.3350684 0.01927952    0.01478492    1.7498662
7 0.02831220 1.3715686 0.01992913    0.01540792    1.4419828
8 0.02849476 1.3949502 0.01810517    0.01505762    0.9997176
9 0.02866079 1.4253270 0.02019306    0.01398133    1.0868125
```

In doing so, we first determined that with R's subset selection function, our data would look more fined with excluding two of our potential factors: coal power generation level in Seoul and the construction expenditure amount in Seoul as by the following:

Figure 3:

```
> fit.full = regsubsets(data$finedust~., data=data, nvmax=11, nbest=3)
경고메시지:
1: In leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax, force.in = force.in,  :
  2  linear dependencies found
2: In leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax, force.in = force.in,  :
  nvmax reduced to  8
> summary(fit.full)
Subset selection object
Call: regsubsets.formula(data$finedust ~ ., data = data, nvmax = 11,
    nbest = 3)
10 Variables  (and intercept)
              Forced in Forced out
car_regist        FALSE      FALSE
bikeroad_lg       FALSE      FALSE
bikeroad_nu       FALSE      FALSE
sub_users         FALSE      FALSE
GRDP              FALSE      FALSE
tree              FALSE      FALSE
greenbelt         FALSE      FALSE
coalpower         FALSE      FALSE
nuclearenergy     FALSE      FALSE
construction      FALSE      FALSE
3 subsets of each size up to 8
Selection Algorithm: exhaustive
         car_regist bikeroad_lg bikeroad_nu sub_users GRDP tree greenbelt
1  ( 1 ) " "        " "         " "         "*"       " "  " "  " "
1  ( 2 ) " "        " "         " "         " "       "*"  " "  " "
1  ( 3 ) " "        " "         " "         " "       " "  " "  "*"
2  ( 1 ) " "        " "         " "         "*"       " "  " "  " "
2  ( 2 ) " "        " "         " "         " "       " "  " "  "*"
2  ( 3 ) " "        " "         " "         "*"       " "  "*"  " "
3  ( 1 ) " "        " "         " "         " "       " "  " "  "*"
3  ( 2 ) " "        " "         " "         "*"       " "  "*"  " "
3  ( 3 ) " "        " "         " "         "*"       " "  " "  "*"
4  ( 1 ) " "        " "         " "         " "       "*"  " "  "*"
4  ( 2 ) " "        "*"         " "         "*"       " "  " "  "*"
4  ( 3 ) " "        " "         " "         "*"       " "  " "  "*"
5  ( 1 ) " "        " "         "*"         "*"       " "  "*"  " "
5  ( 2 ) " "        "*"         " "         " "       " "  "*"  "*"
5  ( 3 ) " "        "*"         " "         "*"       " "  " "  "*"
6  ( 1 ) "*"        "*"         "*"         "*"       " "  " "  "*"
6  ( 2 ) " "        "*"         "*"         "*"       " "  " "  "*"
6  ( 3 ) " "        "*"         " "         "*"       "*"  "*"  " "
7  ( 1 ) " "        "*"         " "         "*"       "*"  "*"  " "
```

After, we have discovered that using 6 of factors to run our regression test will return us a finer plot by running this function in R.

Figure 4:

```
> fit <- lm(data$finedust~data$car_regist+data$bikeroad_lg+data$bikeroad_nu+data$sub_users+data$greenbelt+data$nuclearenergy)
> summary(fit)

Call:
lm(formula = data$finedust ~ data$car_regist + data$bikeroad_lg +
    data$bikeroad_nu + data$sub_users + data$greenbelt + data$nuclearenergy)

Residuals:
       1        2        3        4        5        6        7
-0.011439  0.049317 -0.066126  0.128692 -0.141158  0.164696 -0.145659
       8        9
 0.019987  0.001691

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.108e+02  1.260e+01   8.798 0.012675 *
data$car_regist     2.363e+02  4.399e+01   5.372 0.032944 *
data$bikeroad_lg    2.096e+05  1.713e+04  12.237 0.006612 **
data$bikeroad_nu   -3.115e+05  4.499e+04  -6.923 0.020235 *
data$sub_users     -2.904e-01  8.994e-03 -32.281 0.000958 ***
data$greenbelt     -1.196e+01  8.051e-01 -14.860 0.004498 **
data$nuclearenergy -3.178e+03  2.312e+02 -13.743 0.005253 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2147 on 2 degrees of freedom
Multiple R-squared:  0.9997,    Adjusted R-squared:  0.9989
F-statistic:  1170 on 6 and 2 DF,  p-value: 0.0008543
```

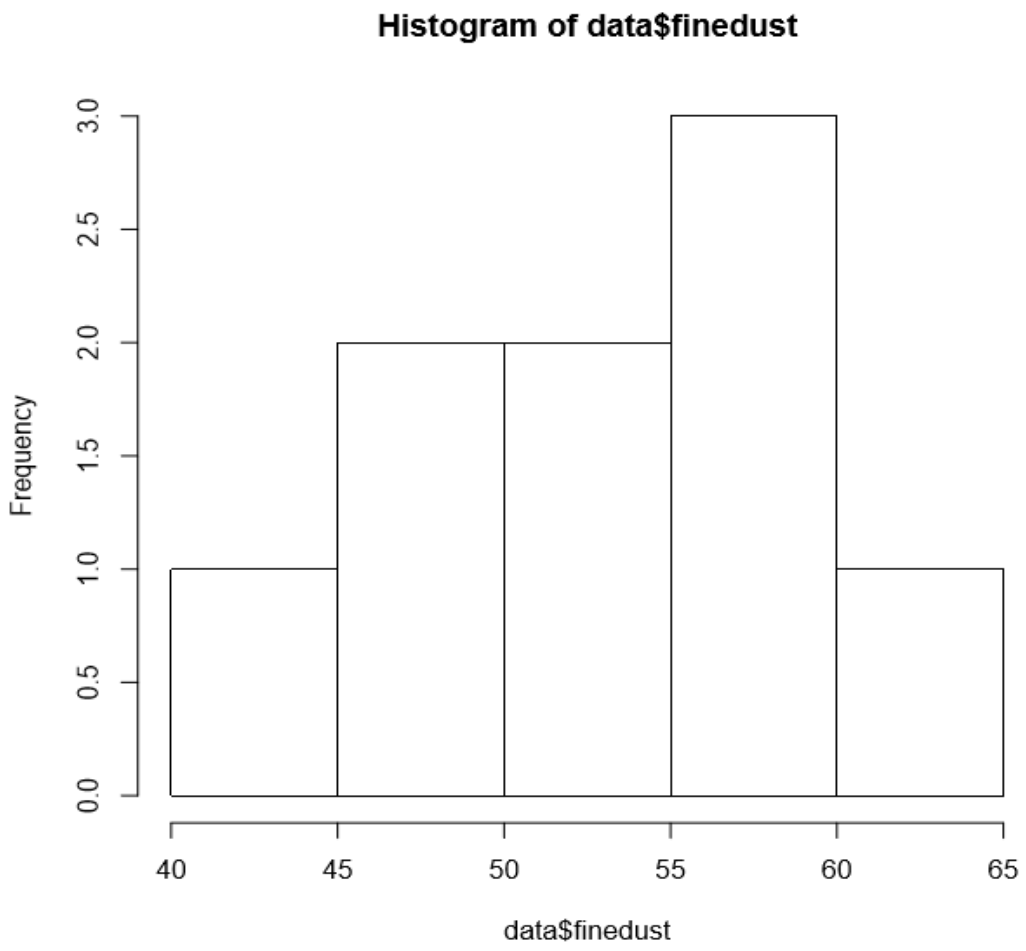
As by the above figure, we be using the potential factors: car_regist, bikeraod_lg, bikeroad_nu, sun_users, greenbelt, and nuclearenergy to better fit our regression test. After, we determined the mean, median, min, max, and standard deviation of each data sets as by the following:

Figure 5:

```
> mean(data$finedust)
[1] 53.22222
> mean(data$car_regist)
[1] 0.2924986
> mean(data$bikeroad_lg)
[1] 7.204444e-05
> mean(data$bikeroad_nu)
[1] 3.681489e-05
> mean(data$sub_users)
[1] 238.7106
> mean(data$greenbelt)
[1] 1.192359
> mean(data$nuclearenergy)
[1] 0.01472767
> min(data$finedust)
[1] 43
> max(data$finedust)
[1] 61
> min(data$car_regist)
[1] 0.2805594
> max(data$car_regist)
[1] 0.2995932
> min(data$bikeroad_lg)
[1] 6.28e-05
> max(data$bikeroad_lg)
[1] 8.4e-05
> min(data$bikeroad_nu)
[1] 2.97e-05
> max(data$bikeroad_nu)
[1] 4.22e-05
> min(data$sub_users)
[1] 225.817
> max(data$sub_users)
[1] 263.8956
> min(data$greenbelt)
[1] 0.7710894
> max(data$greenbelt)
[1] 1.425327
> min(data$nuclearenergy)
[1] 0.01398133
> max(data$nuclearenergy)
[1] 0.01540792
> median(data$finedust)
[1] 55
> median(data$car_regist)
[1] 0.2925509
> median(data$bikeroad_lg)
[1] 7.12e-05
> median(data$bikeroad_nu)
[1] 3.68e-05
> median(data$sub_users)
[1] 227.5413
> median(data$greenbelt)
[1] 1.298028
> median(data$nuclearenergy)
[1] 0.01478492
> sd(data$finedust)
[1] 6.359595
> sd(data$car_regist)
[1] 0.006241698
> sd(data$bikeroad_lg)
[1] 6.994303e-06
> sd(data$bikeroad_nu)
[1] 4.372447e-06
> sd(data$sub_users)
[1] 15.11671
> sd(data$greenbelt)
[1] 0.2470944
> sd(data$nuclearenergy)
[1] 0.0004272958
```

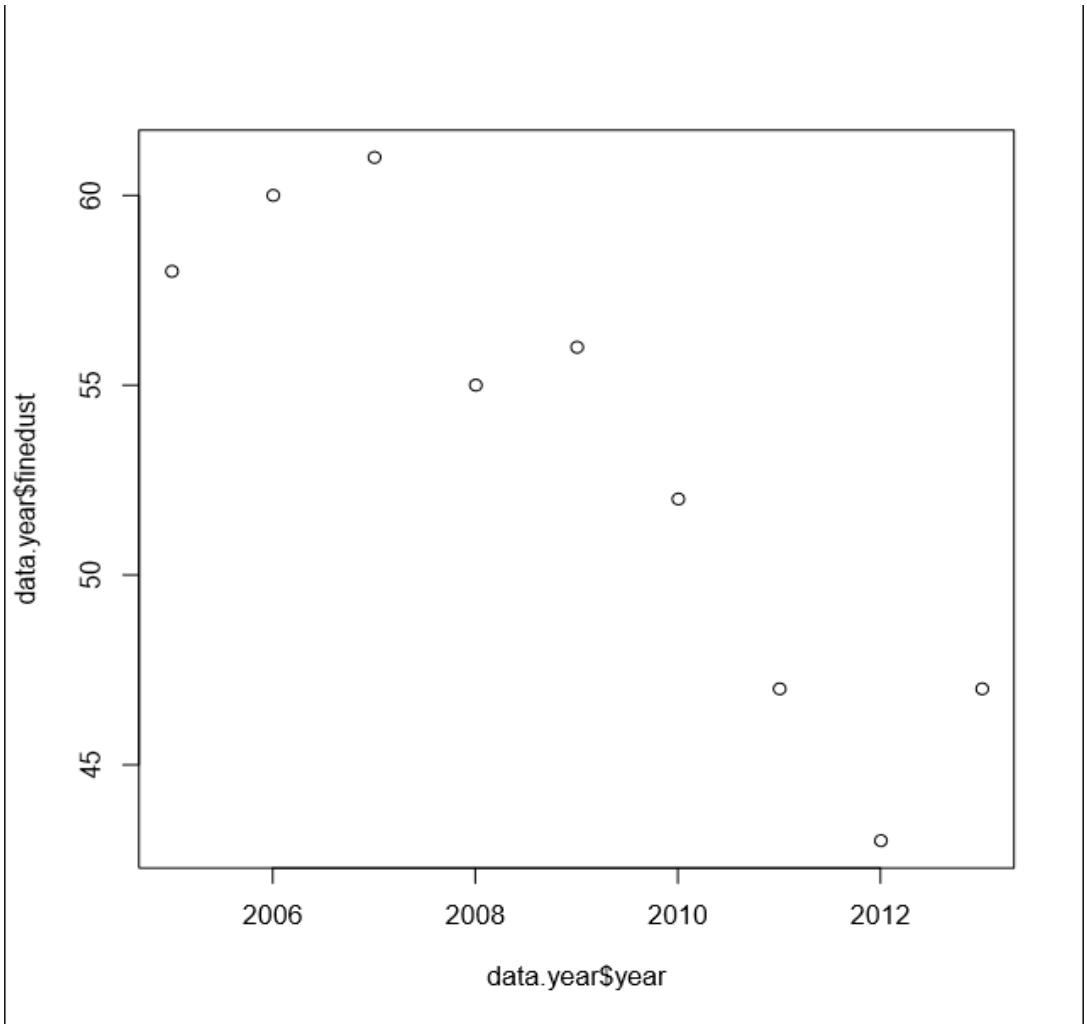
As by the following figure, we have discovered the corresponding values as such and will continue to look for significance in the next steps.

Figure 6:



Above figure is the histogram of fine dust level in Seoul from 2005 to 2013. The units are in $\mu g/m^3$ and ranges from 40 to 65. The above histogram displays the min to max of the various fine dust levels in Seoul. Accordingly below is the plot of the fine dust levels in Seoul from 2005 to 2013.

Figure 7:



As from our findings, we would like to do the following steps in the next steps to better establish our findings:

1) Discover why the subset selection function returned to us to exclude other potential factors.
2) Discover why the subset selection function returned to us to include the listed factors. For what reason?
3) Label the plots and graphs that we will use in our final presentation.
4) Run a t.test to test our null hypothesis
5) Run a regression test to see why certain factors were more highly correlated than the other factors. Same with the one with the least correlation.
6) Do research after finding the effects of the regression and the regression values to implement why out findings were as is.
7) Present an amazing presentation to class.