Section 1. Statistical Test

1.1 Which statistical test did you use to analyze the NYC subway data? Did you use a one-tail or a two-tail P value? What is the null hypothesis? What is your p-critical value?

a) The statistical test used to analyze the NYC subway data is Mann Whitney U-test.
b) I used two- tail P value.
c) H- Null: Probability that number of people entering the subway during no rain is equal as the probability that number of people entering the subway during rain.
d) p-critical : 0.05 (alpha – 0.05 for two-tail)

1.2 Why is this statistical test applicable to the dataset? In particular, consider the assumptions that the test is making about the distribution of ridership in the two samples.

Assumption:

Assumption #1: The dependent variable should be measured at the ordinal or continuous level.

Assumption #2: The independent variable should consist of two categorical, independent groups

Assumption #3: There should be no relationship between the observations in each group or between the groups themselves.

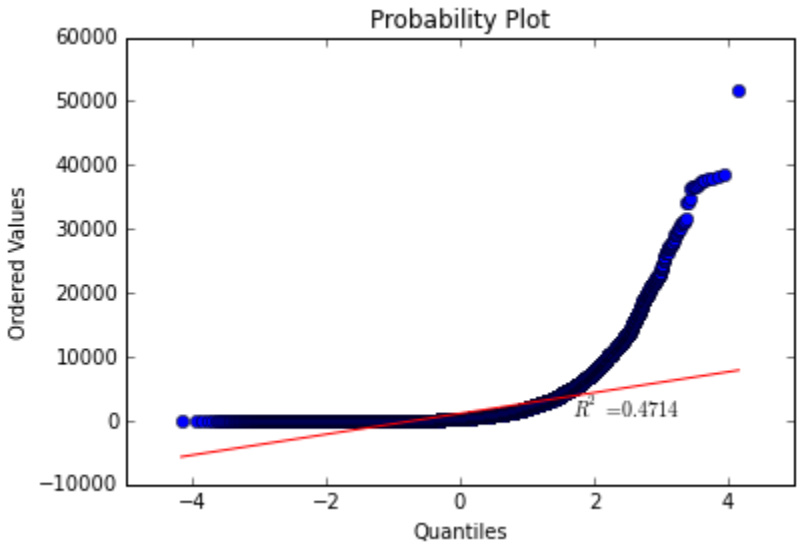
Assumption #4: The dependent variable is not normally distributed and their distributions should also have the same shape.

Proof:

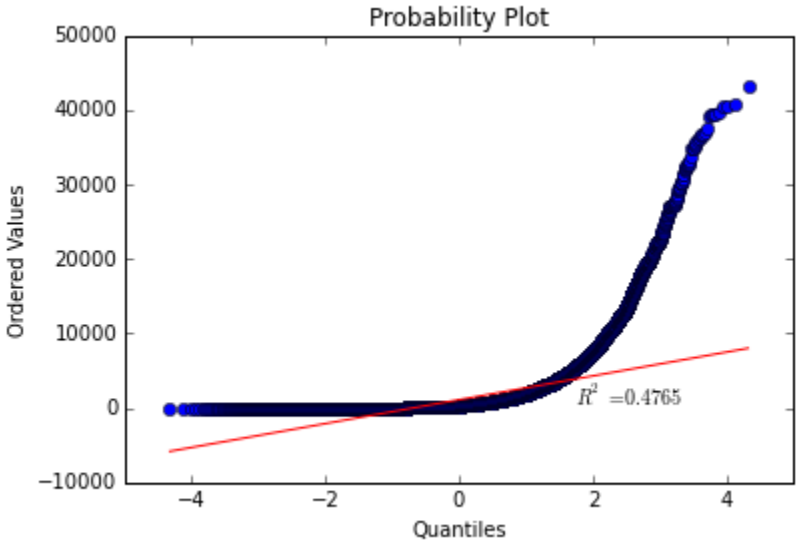
- Independent variable or variable that has changes to influence the value of our outcome variable = **rain and no_rain.**
- Dependent variable or the variable that we consider our outcome = **number of entries.**

- Assumption 1: The number of people entering the subway can be ordered from smallest to maximum entry. Those the dependent variable is measured at the ordinal level for original data set.
- The independent variable is a two categorical group rain and no rain since the values in the group are discrete and it is just a labor and has no relationship or sequence.
- The two group rain and no rain are independent even so there is no relationship between a person taking subway during sunny day is going to affect the probability of a person not taking the subway during rainy days. This is because I thing rain is a natural event so it must be random event.
- The dependent variable definitely violate normality. This can be clearly seen from the qq-plot

Rainy day



No rain



1.3 What results did you get from this statistical test? These should include the following numerical values: p-values, as well as the means for each of the two samples under test.

U- Value =1024409167

p-value= .038

(Mean for rainy day, mean for no rain day) = (1105, 1090)

(median of rainy day, median foe no rain day) = (282,278)

1.4 What is the significance and interpretation of these results?

The p-value is lesser than 0.05 so there is a significance difference so we reject the null hypotheses. Thus the probability of the number of people entering the subway during the rainy days is not same as the probability of number of people entering the subway during the sunny days.

Section 2. Linear Regression

2.1 What approach did you use to compute the coefficients theta and produce prediction for ENTRIESn_hourly in your regression model?

I used OLS using Statsmodels to compute the coefficients theta and produce prediction for ENTRIESm_hourly in your regression model.

2.2 What features (input variables) did you use in your model? Did you use any dummy variables as part of your features?

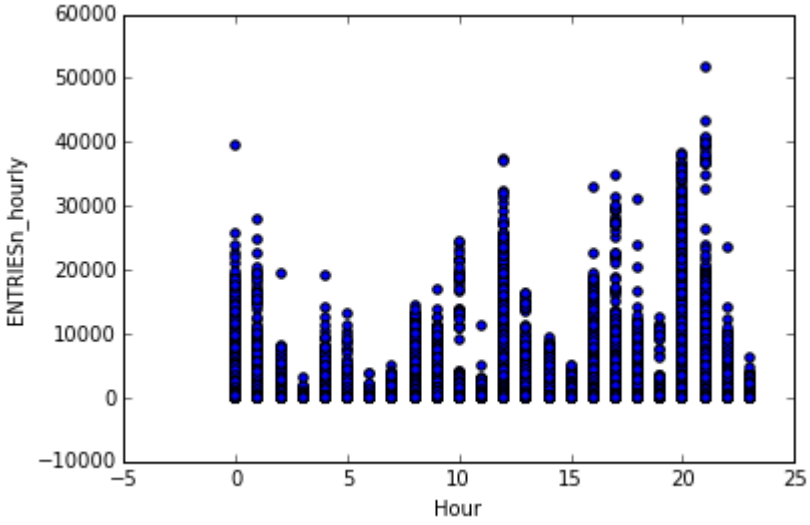
The feature I used in my model are as follow units (station), Hours, maxtemp, fog, rain, and unit. Yes I did use a dummy variable “unit (station)” fog, rain, hours.

2.3 Why did you select these features in your model? We are looking for specific reasons that lead you to believe that the selected features will contribute to the predictive power of your model.

The reasons for the inclusion of these feature are as follow:

UNIT (station_no)= the reason behind using the unit as a dummy feature is based on data exploration and experimentation as soon I included it in my model, it drastically improved the model.

Hour = the reason I think hour is one of the feature that must be included in the production is that depending upon the hour the ridership will change. There ridership is more at peak hours and less at other times. The R^2 also increase by .3 from .41 to .45. Look at the scatter plot you can see the variation



Fog: I thought that when it is foggy people don't want to use their private transport so I thought of including the fog to the model. There was an increase in the R^2.

Maxtemp: I think if the climate is very hot people for comfort reason may opt not to travel travel in crowds and might choses transport such as private cars. Thus decreasing the number of entry_hour. That's why I think I should add maxtemp temperature.

2.4 What are the parameters (also known as "coefficients" or "weights") of the non-dummy features in your linear regression model?

Maxtempi= -34

Which means for every one unit rise in temp there is drop of 34 number of people enter_hour.

OLS Regression Results

Dep. Variable:	ENTRIESn_hourly	R-squared:	0.501
Model:	OLS	Adj. R-squared:	0.499
Method:	Least Squares	F-statistic:	270.2
Date:	Tue, 22 Sep 2015	Prob (F-statistic):	0.00
Time:	15:32:45	Log-Likelihood:	-1.1648e+06
No. Observations:	131951	AIC:	2.331e+06
Df Residuals:	131461	BIC:	2.335e+06
Df Model:	489		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
const	1095.3485	4.552	240.630	0.000	1086.427	1104.270
fog	29.0618	4.554	6.382	0.000	20.136	37.988
maxtempi	-34.5023	4.554	-7.576	0.000	-43.428	-25.576
unit_R002	-114.2444	6.407	-17.831	0.000	-126.802	-101.687

2.5 What is your model's R2 (coefficients of determination) value?

The R^2 value is 0.501.

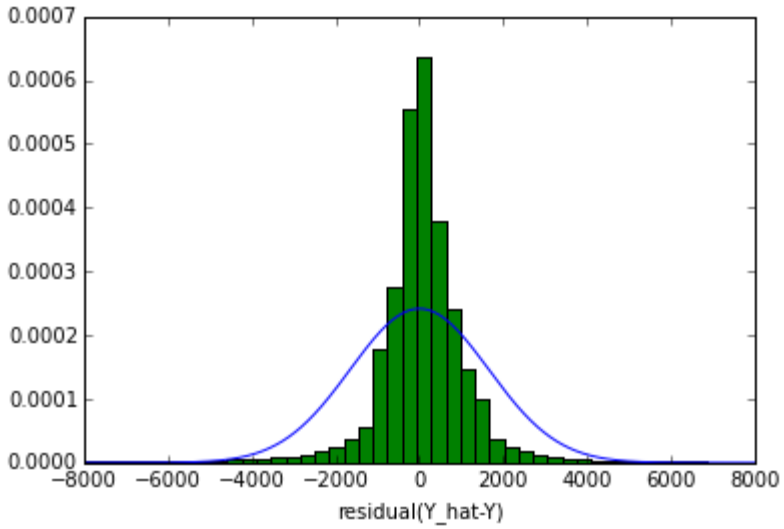
OLS Regression Results

Dep. Variable:	ENTRIESn_hourly	R-squared:	0.501
Model:	OLS	Adj. R-squared:	0.499
Method:	Least Squares	F-statistic:	270.2
Date:	Tue, 22 Sep 2015	Prob (F-statistic):	0.00

2.6 What does this R2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R2 value?

The R^2 value explains, what percentage of total variance in the model is described by our feature. The R^2 for our model is .50. Thus 50% of variance in the model is explained by the model and the rest is explained by some other unknown feature.

Residuals histogram for hourly entry

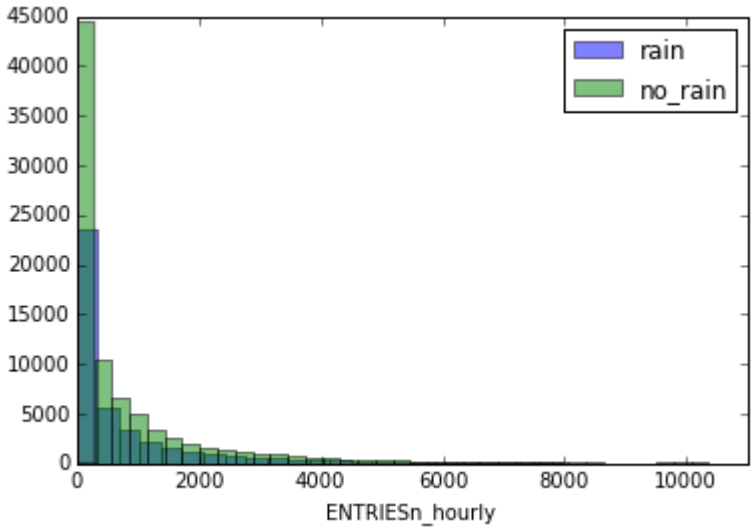


The residual histogram is in the shape of bell-curve thus the model follows the normal-distribution and our model is said to be good fit.

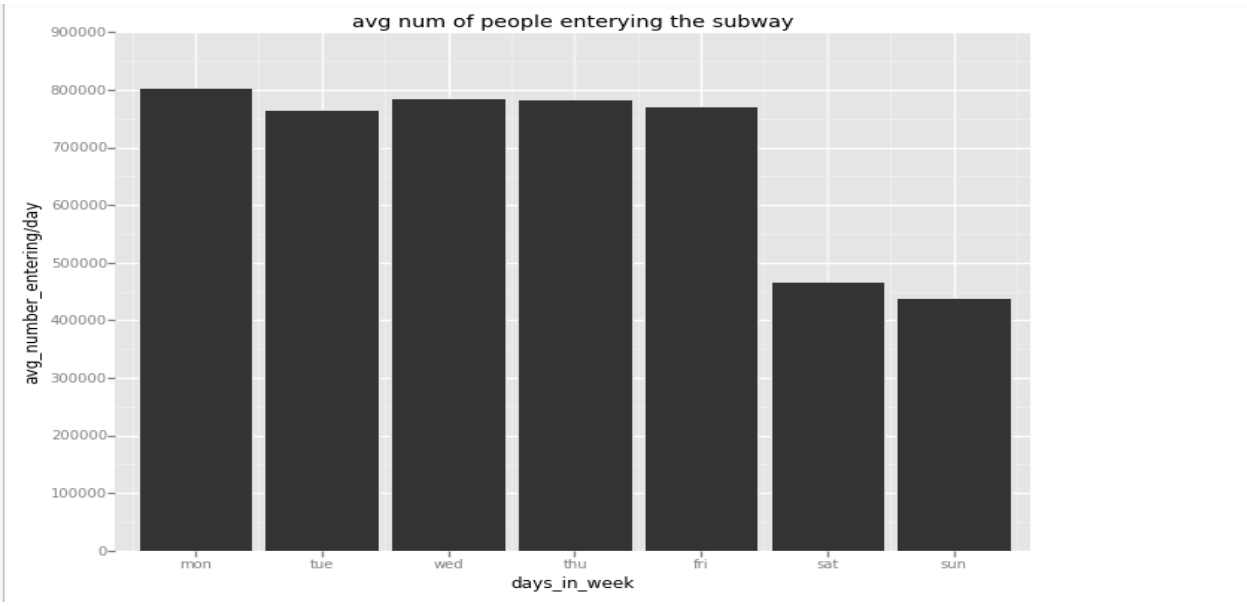
SECTION 3:

3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.

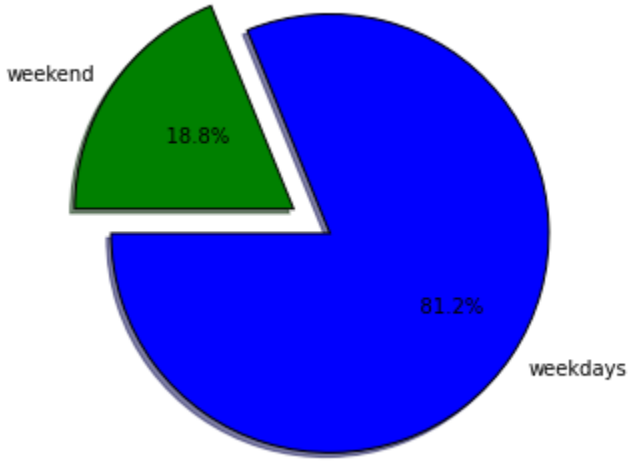
Histogram to find the frequency of entries_hours during rain and no_rain



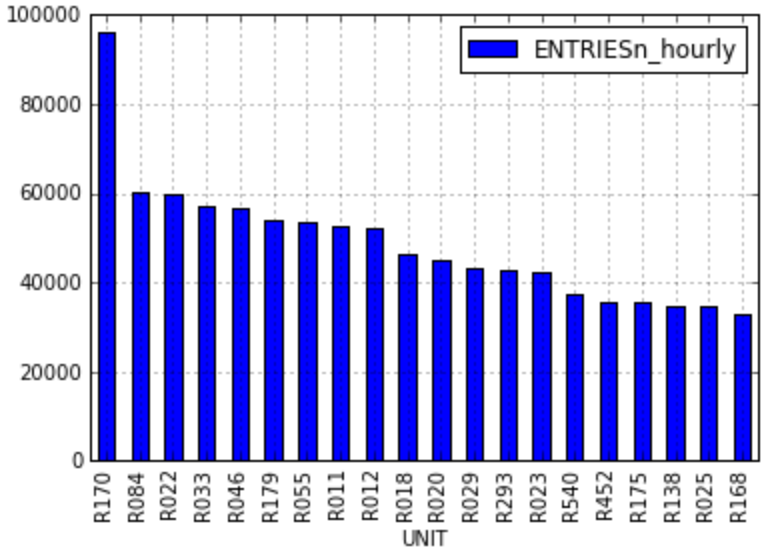
3.2 One visualization can be more freeform. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots) or attempt to implement something more advanced if you'd like. Some suggestions are:



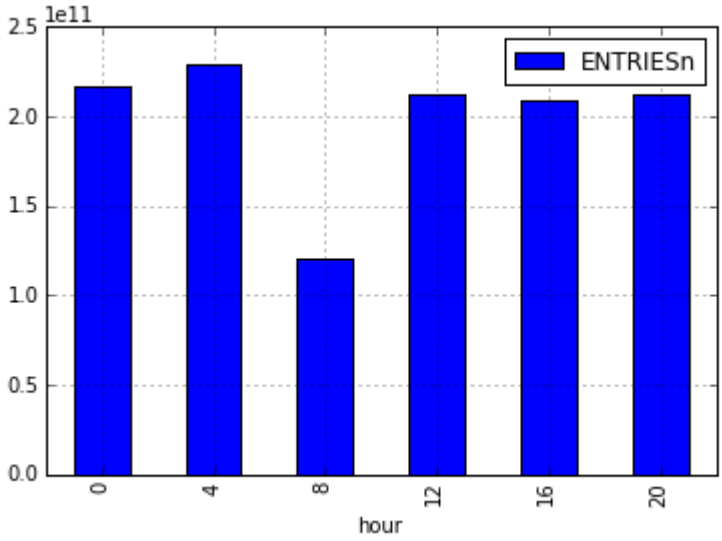
Bar graph to show avg_number_entry in days in a week



Pie chart to show the percentage of ridership during weekend and weekdays.



Bar chart which rains the station based on the highest number of ridership



Ridership based on the hours

Section 4. Conclusion

Please address the following questions in detail. Your answers should be 1-2 paragraphs long.

4.1 From your analysis and interpretation of the data, do more people ride
the NYC subway when it is raining or when it is not raining?

No there is not much difference in the ridership during rainy days and sunny days.

4.2 What analyses lead you to this conclusion? You should use results from both your statistical

Tests and your linear regression to support your analysis.

Statistical Tests:

The p-value is lesser than 0.05 so there is a significance differences so we reject the null hypotheses. Thus the probability of the number of people entering the subway during the rainy days is not same as the probability of number of people entering the subway during the sunny days

The mean for number of entery_hour in the rainy days is 1105 compared to number of entery_hour during sunny day is 1090.

Thus people using the subway during the rainy day is more than the sunny day but if you see the difference it's close to 15 people travel more in the rainy days. In practically 15 people is not a big crowed when compared to the volume of the NY subway users.

Regression analysis:

The following reason are the validation for my conclusion:

The feature 'rain' has a lager P-value and it has no significant difference thus we accept null hypothesis which says the coefficient is equal to zero and has no effect to the prediction model. The confident interval has 0 in its range (-14, 5) which also means we accept null hypothesis.

Regression test with rain:

OLS Regression Results

Dep. Variable:	ENTRIESn_hourly	R-squared:	0.501
Model:	OLS	Adj. R-squared:	0.499
Method:	Least Squares	F-statistic:	270.0
Date:	Mon, 21 Sep 2015	Prob (F-statistic):	0.00
Time:	23:07:27	Log-Likelihood:	-1.1649e+06
No. Observations:	131951	AIC:	2.331e+06
Df Residuals:	131461	BIC:	2.335e+06
Df Model:	489		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
const	1095.3485	4.553	240.578	0.000	1086.425	1104.272
fog	30.9869	5.086	6.093	0.000	21.019	40.955
rain	-4.4201	5.086	-0.869	0.385	-14.388	5.548

What do the coefficients tell us about the features?

	coef	std err	t	P>|t|	[95.0% Conf. Int.]	
const	1095.3485	4.552	240.630	0.000	1086.427	1104.270
fog	29.0618	4.554	6.382	0.000	20.136	37.988
maxtempi	-34.5023	4.554	-7.576	0.000	-43.428	-25.576
unit_R002	-114.2444	6.407	-17.831	0.000	-126.802	-101.687
unit_R003	-149.7611	6.346	-23.600	0.000	-162.199	-137.324
unit_R004	-139.4860	6.382	-21.856	0.000	-151.995	-126.977

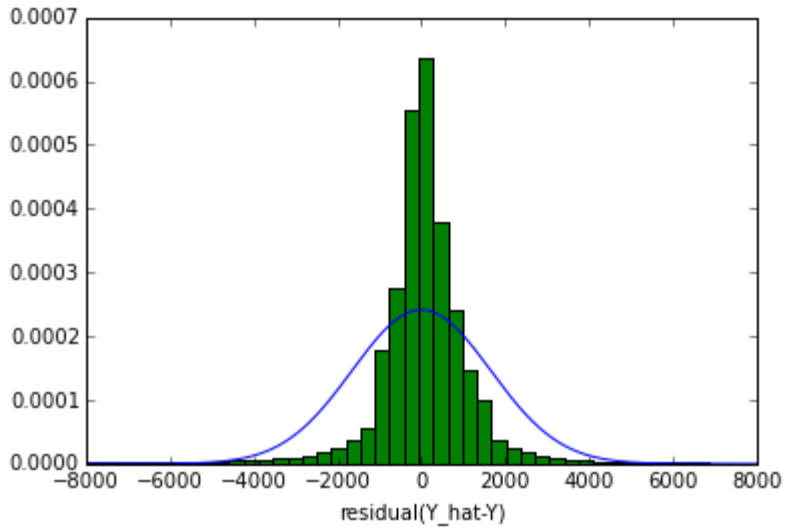
The equation shows that the coefficient for maxtempi is -34 which means that for every one temperature rise there is a fall of 34 number of people in entry_hour.

If Dummy-variables are used, how are their coefficients compared to the weather variables' coefficients?

The dummy variable are interpreted as in case of fog there are two dummy variable 0 for no fog and 1 for fog day. I have in included constant in my model and dropped the first variable of all the dummy variable to avoid dummy trap. Thus the constant become the coefficient of the first variable in the case of fog the coefficient of the no fog day is 1095 and the coefficient of the fog days are 1095+29= 1126. Thus the feature fog day add 1126number of people to entery_hour during foggy day and feature no_fog adds 1095 people to entery_hour during fog days. This is how the feature unit,hour dummy variable are also interpreted.

How valid is our Regression analysis?

Residuals histogram for hourly entry



The residual histogram is in the shape of bell-curve thus the model follows the normal-distribution and our model is said to be good fit.

5.1 Please discuss potential shortcomings of the methods of your analysis, including:

1. Dataset,

   I think the collection of data was not uniform. For example the hour at with the number of entery, was not measured at the same interval for every station. The improved data had ratified this defect but it has to drop a lot to data to archive this for instance the station R170 which had the highest traffic has to be dropped. Dropping data will always affect the prediction in our model.

2. Analysis, such as the linear regression model or statistical test.

   The dependent variable(entriy_hour) from the data-set has many zero values and my regression model is the line as my best fit in cases where the entries are around zero the prediction number will have negative number which make no sense to have number of people in negative. This can be overcome by using different regression method such as logistic regression or doing transformation.

Regression test without rain feature:

OLS Regression Results

Dep. Variable:	ENTRIESn_hourly	R-squared:	0.501
Model:	OLS	Adj. R-squared:	0.499
Method:	Least Squares	F-statistic:	270.5
Date:	Tue, 22 Sep 2015	Prob (F-statistic):	0.00
Time:	00:01:18	Log-Likelihood:	-1.1649e+06
No. Observations:	131951	AIC:	2.331e+06
Df Residuals:	131462	BIC:	2.335e+06
Df Model:	488		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
const	1095.3485	4.553	240.578	0.000	1086.425	1104.272
fog	29.0206	4.555	6.371	0.000	20.093	37.948
unit_R002	-114.2316	6.409	-17.825	0.000	-126.792	-101.671
unit_R003	-149.6520	6.347	-23.578	0.000	-162.092	-137.212