Health Insurance claim

CAUSE AND EFFECT ANALYSIS

-APEKSHAA.V

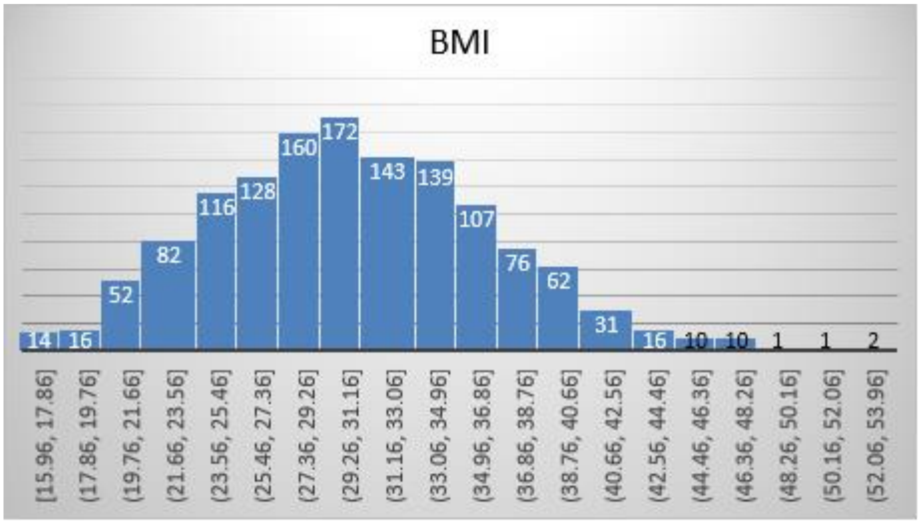
1) Perform the Exploratory Data Analysis on the data.

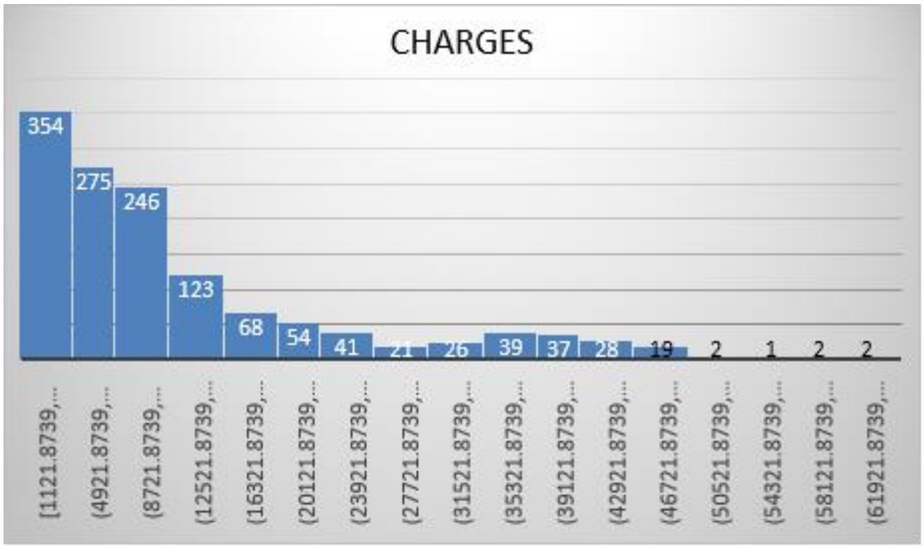
a) Identify the categorical and continuous variables

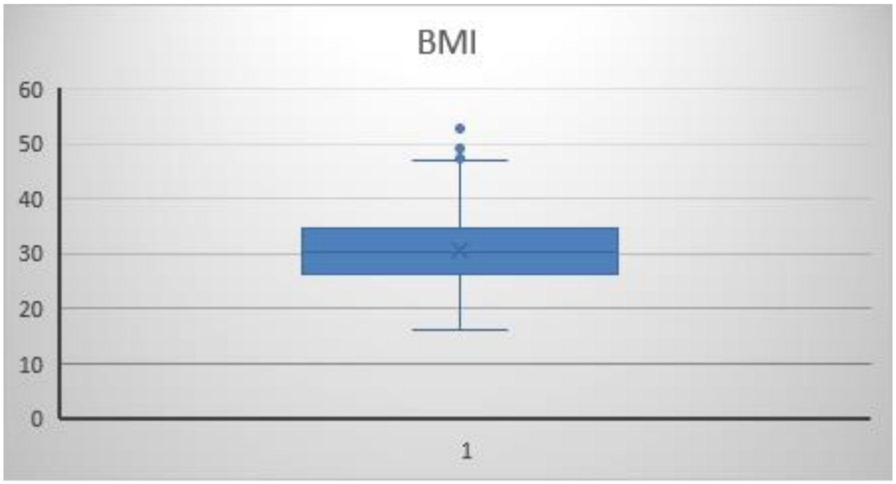
Categorical Variables	Continious Variables
Sex	Bmi
Smoker	Charges
Region	

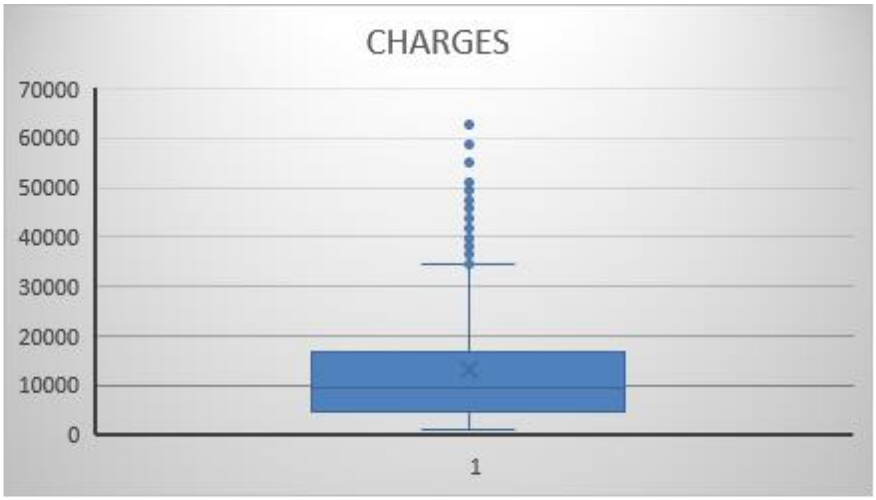
Age and **Children** comes under the category of discrete.

b) Make Histograms and box plots (univariate analysis) for continuous variables and do a correlation analysis (multivariate analysis)









CORRELATION ANALYSIS

	bmi	**charges($)**
bmi	1	
charges($)	0.198340969	1

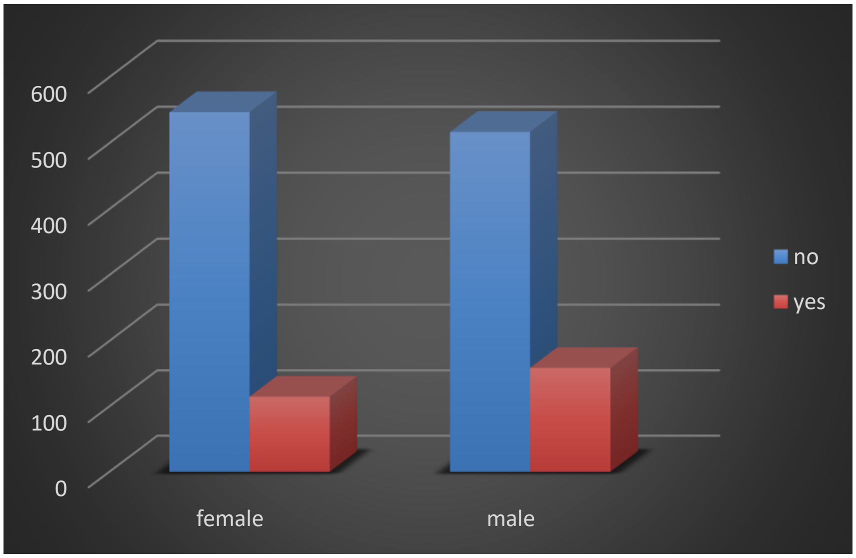
c) Make relevant Pivot tables and charts for:

i. Male/Female ratio and share information on which gender has more smokers

Count of smoker	Column Labels	
Yes/No	**female**	**male**
no	547	517
yes	115	159
Grand Total	**662**	**676**

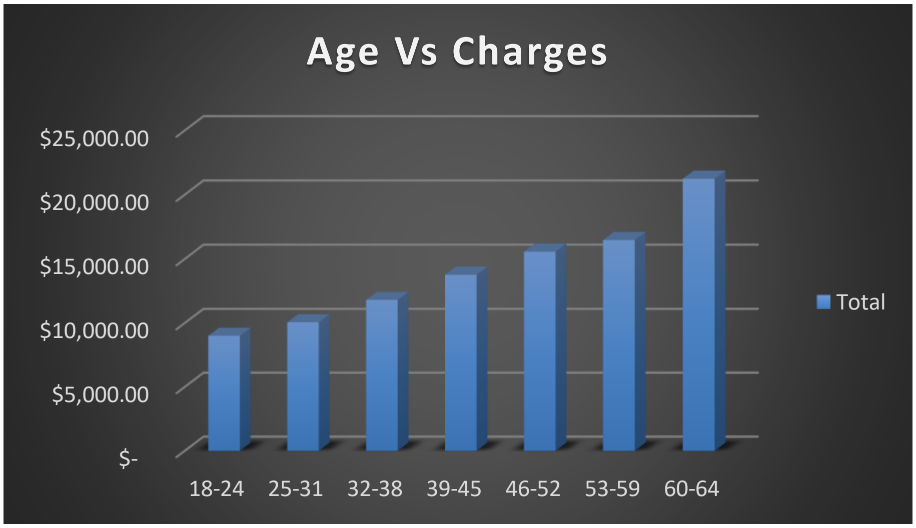
Male by Female Ratio = **1.382608696**

Since the male to female ratio is above unity (1), we understand that the make has more smokers.



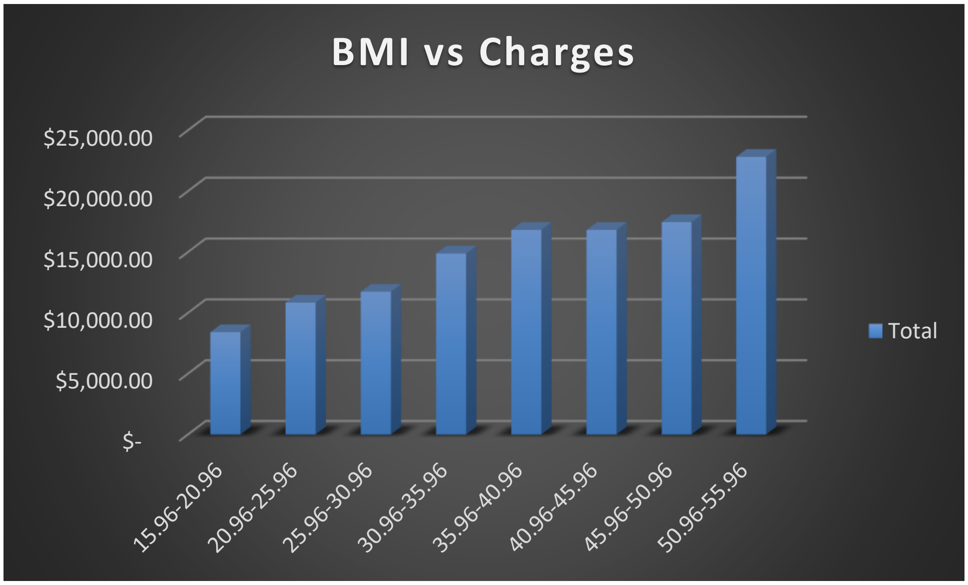
ii. Charges vs Age

Age	Average of charges($)
18-24	$ 9,011.34
25-31	$ 10,065.69
32-38	$ 11,818.41
39-45	$ 13,778.32
46-52	$ 15,575.13
53-59	$ 16,476.98
60-64	$ 21,248.02



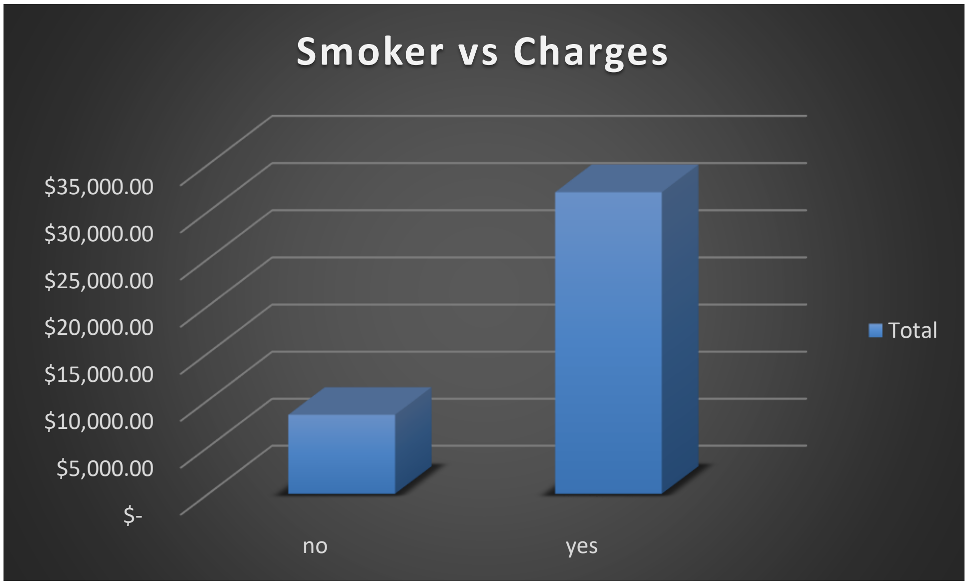
iii. Charges vs BMI

BMI	Average of charges($)
15.96-20.96	$ 8,427.01
20.96-25.96	$ 10,859.07
25.96-30.96	$ 11,756.59
30.96-35.96	$ 14,891.44
35.96-40.96	$ 16,833.61
40.96-45.96	$ 16,829.56
45.96-50.96	$ 17,468.71
50.96-55.96	$ 22,832.43



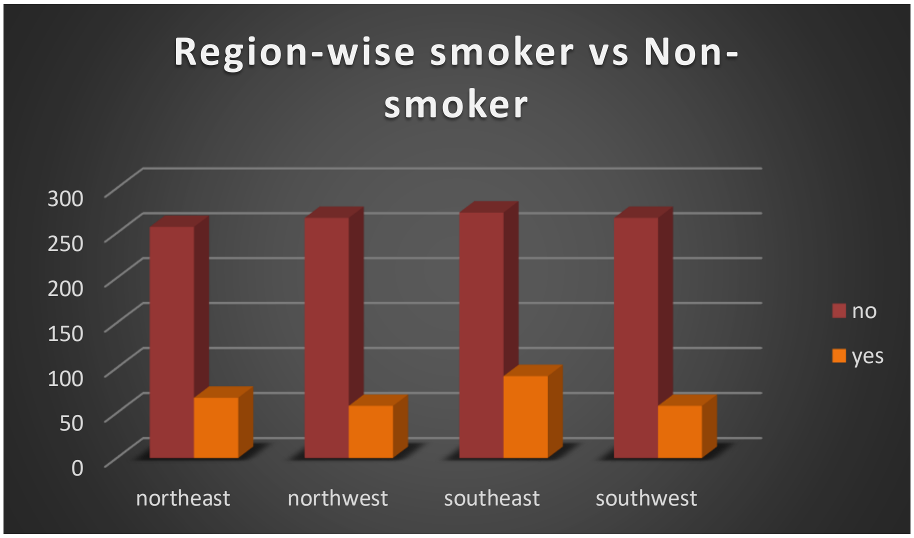
iv. Charges for Smokers vs Non-smokers

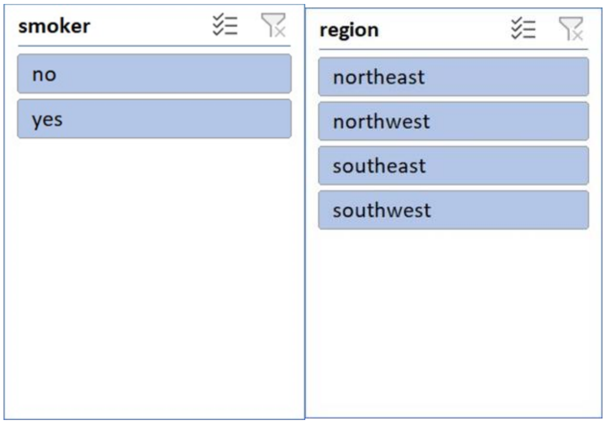
Smoker	Average of charges($)
no	$ 8,434.27
yes	$ 32,050.23



d) Region-wise smokers vs Non-smokers analysis with one or more pivot table and charts

Count of smoker	Column Labels	
Region	**no**	**yes**
northeast	257	67
northwest	267	58
southeast	273	91
southwest	267	58




smoker

no

yes

region

northeast

northwest

southeast

southwest

e) Region-wise charges for smokers vs non-smokers

Average of charges($)	Column Labels	
Region	no	yes
northeast	9165.531672	29673.53647
northwest	8556.463715	30192.00318
southeast	8032.216309	34844.99682
southwest	8019.284513	32269.06349

f) Has charges got something to do with the number of dependents ?

The Correlation between number of dependents and charges = 0.067998

The number of dependents and the charges are directly proportional. So, if the no of dependents are increased, the charges are also increased.

g) Do a similar dependants-charges analysis, Region-wise

Charges($).	Dependents					
Region	**0**	**1**	**2**	**3**	**4**	**5**
northeast	11626.5	16310.2	13615.2	14409.9	14485.2	6979.0
northwest	11324.4	10230.3	13464.3	17786.2	11347.0	8965.8
southeast	14309.9	13687.0	15728.5	18449.8	14451.0	10115.4
southwest	11938.5	10406.5	17483.5	10402.4	14933.3	8444.2

h) Do at least one more pivot table and chart of your own choice on the remaining variables

Count of bmi	Column Labels	
Region	**female**	**male**
northeast	161	163
northwest	164	161
southeast	175	189
southwest	162	163

i) Give your understanding from the patterns observed in point (b)

Interpretation for observations made in point (b)

- The BMI and the charges are the univariate variables that are normally distributed.
- The BMI of 1st quartile is 26.2 and 3rd quartile us 37.4
- The data in charges are positively skewed

j) Give your interpretation for observations made in point (c)

Interpretation for observations made in point (c)

- Males has increased number of smokers.
- The BMI range of 45-50 has highest average charge of 17547.92675.
- Average charges for smokers are four times the charges for non-smokers.
- The Age group 55-65 has the highest average charge of 18513.26.

2) Edit the data as following, to obtain dummy variables:

a) Sex : Replace all the “Males” with “1” and “Females” with “0”, creating numerical entries for gender this way will help you do analysis further. You can use the “Replace with Match entire cell content” option. Do a replace all to save time.

b) Smoker: Replace all the “Smokers” with “1” and “Non-smokers” with “0”.

c) Region: We always create one less category column for the dummy data w.r.t the categories available for that original variable. So for Region, we will create three dummy columns, assuming “Northeast” as zero and omit the column for it. Now create three columns for “northwest”, “Southeast”, “Southwest”. Whichever row has “northwest” region as an entry will take “1” as an entry otherwise “0” in “northwest” column. Similarly in the “Southeast” column, whichever row had “southeast” as an entry will take “1” as the new entry and “0” for the rest of the column (Southeast). Do a similar operation on the “Southwest” column. Please refer to the below image for your understanding,

I used Find and Replace function to make the changes accordingly

sex	smoker	southwest	northwest	southeast
0	1	1	0	0
1	0	0	0	1
1	0	0	0	1
1	0	0	1	0
1	0	0	1	0
0	0	0	0	1
0	0	0	0	1
0	0	0	1	0
1	0	0	0	0
0	0	0	1	0
1	0	0	0	0
0	1	0	0	1
1	0	1	0	0
0	0	0	0	1
1	1	0	0	1
1	0	1	0	0
0	0	0	0	0
1	0	0	0	0
1	0	1	0	0
1	1	1	0	0

3) Do a descriptive summary analysis for the edited data. Perform a Multiple Linear Regression analysis to identify which variables decide the insurance charges/billed insurance claim. Give your interpretation for the above analysis, do another set of regression analysis by dropping insignificant variables, if needed.

Descriptive Summary Analysis of edited data

We use the summary statistics in the data analytics function

Summary	***Age***	***BMI***	***Children***	***Sex***	***Smoker***	***Southwest***	***Northwest***	***Southeast***	***Charges***
Mean	39.2	30.7	1.1	0.5	0.2	0.2	0.2	0.3	13270
Standard Error	0.4	0.2	0.0	0.0	0.0	0.0	0.0	0.0	331
Median	39.0	30.4	1.0	1.0	0.0	0.0	0.0	0.0	9382
Mode	18.0	32.3	0.0	1.0	0.0	0.0	0.0	0.0	1640
Standard Deviation	14.0	6.1	1.2	0.5	0.4	0.4	0.4	0.4	12110
Sample Variance	197.4	37.2	1.5	0.3	0.2	0.2	0.2	0.2	146652372
Kurtosis	-1.2	-0.1	0.2	-2.0	0.1	-0.6	-0.6	-0.9	2
Skewness	0.1	0.3	0.9	0.0	1.5	1.2	1.2	1.0	2
Range	46.0	37.2	5.0	1.0	1.0	1.0	1.0	1.0	62649
Minimum	18.0	16.0	0.0	0.0	0.0	0.0	0.0	0.0	1122
Maximum	64.0	53.1	5.0	1.0	1.0	1.0	1.0	1.0	63770
Sum	52459.0	41027.6	1465.0	676.0	274.0	325.0	325.0	364.0	17755825
Count	1338.0	1338.0	1338.0	1338.0	1338.0	1338.0	1338.0	1338.0	1338

We use the regression analysis in data analytics function from the data tab for Multiple Linear Regression analysis

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.866552384
R Square	0.750913035
Adjusted R Square	0.74941364
Standard Error	6062.102289
Observations	1338

	df	*SS*	*MS*	*F*	*Significance F*
Regression	8	1.47235E+11	1.8404E+10	500.8107	0
Residual	1329	48839532844	36749084.2		
Total	1337	1.96074E+11			

	Coefficients	***Standard Error***	***t Stat***	***P-value***	***Lower 95%***	***Upper 95%***	***Lower 95.0%***	***Upper 95.0%***
Intercept	-11938.5	987.8	-12.1	0.0	-13876.4	-10000.7	-13876.4	-10000.7
age	256.9	11.9	21.6	0.0	233.5	280.2	233.5	280.2
bmi	339.2	28.6	11.9	0.0	283.1	395.3	283.1	395.3
children	475.5	137.8	3.5	0.0	205.2	745.8	205.2	745.8
sex	-131.3	332.9	-0.4	0.7	-784.5	521.8	-784.5	521.8
smoker	23848.5	413.2	57.7	0.0	23038.0	24659.0	23038.0	24659.0
southwest	-960.1	477.9	-2.0	0.0	-1897.6	-22.5	-1897.6	-22.5
northwest	-353.0	476.3	-0.7	0.5	-1287.3	581.4	-1287.3	581.4
Southeast	-1035.0	478.7	-2.2	0.0	-1974.1	-95.9	-1974.1	-95.9

AVERAGE = 42.0353%

ACCURACY = 57.9647%

Interpretation for the above analysis

- **From this analysis we can observe that the insignificant variables is sex**
- **The variable Smokers have a p value, i.e it is the most significant variable.**
- **This model has a accuracy of 57.964%.**

Observing p-value

Model created after removing the variable **sex**

Regression Statistics	
Multiple R	0.8145016
R Square	0.6634129
Adjusted R Square	0.6618956
Standard Error	7041.5777
Observations	1338

	df	*SS*	*MS*	*F*	*Significance F*
Regression	6	1.3E+11	2.17E+10	437.2333	0
Residual	1331	6.6E+10	49583817		
Total	1337	1.96E+11			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-4066.9	1054.5	-3.9	0.0	-6135.4	-1998.3	-6135.4	-1998.3
bmi	410.1	33.0	12.4	0.0	345.4	474.8	345.4	474.8
children	595.4	159.9	3.7	0.0	281.7	909.1	281.7	909.1
smoker	23629.7	478.4	49.4	0.0	22691.2	24568.3	22691.2	24568.3
southwest	-1030.4	555.1	-1.9	0.1	-2119.4	58.7	-2119.4	58.7
northwest	-390.6	553.2	-0.7	0.5	-1475.8	694.7	-1475.8	694.7
southeast	-1409.1	555.7	-2.5	0.0	-2499.2	-319.0	-2499.2	-319.0