NYU TANDON SCHOOL OF ENGINEERING

Project II

Business Analysis _ MG-GY 9753

Author:

Menghe Dou

Yichun Zhou

Yichi Zhang

Professor:

Dr. Gazi Md Daud Iqbal

March, 2020

We *Menghe Dou, Yichun Zhou, Yichi Zhang* did not give or receive any assistance on this project from other groups, and the report submitted is wholly by us.

Signature: Yichi Zhang Yichun Zhou

Menghe Dou

Proposal

- ***Problem Statement***

Data Set (attached in Appendices I):

Data of house price from Kaggle dataset (https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data), which is a subsidiary of Google LLC and a website contained open datasets for machine learner.

We randomly chose 100 sets of data and used some columns as variables to predict the price of house.

Variables:

SalePrice – respond variable.

LotFrontage – predictor variable, which describes linear feet of street connected to property

LotArea – predictor variable, lot size in square feet

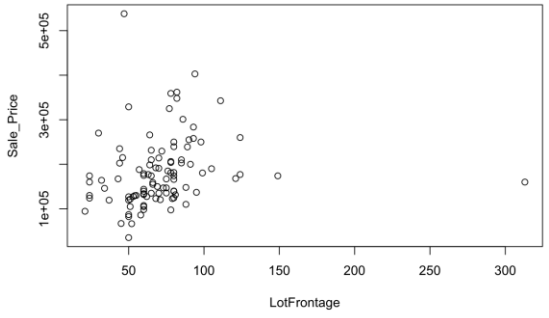
GrLivArea – predictor variable, indicate above grade (ground) living area square feet

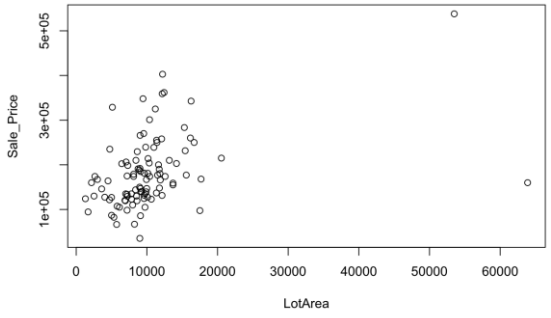
Total rooms – predictor variable, indicate the number of total bathrooms and bedrooms

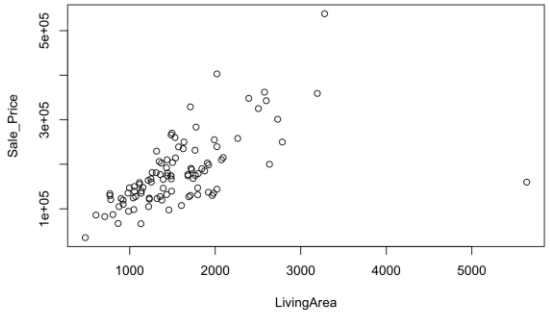
- ***Meaning***

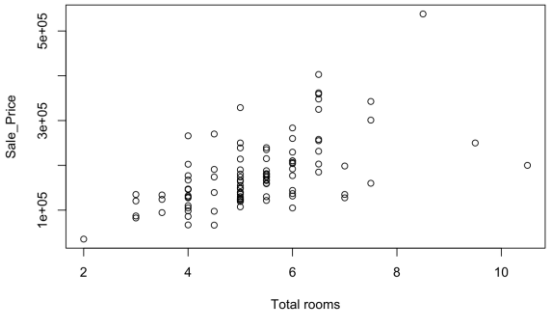
Housing market has always been the hot topic in our life. By figuring out what kind of factor will affect price the most, buyers can make decision more reasonable and effectively and house owner can predict their house price more precisely. (What directly affects the house price is often the house structure, area, geographical location and other properties of the house itself.) We assume that lot frontage, lot area, ground living area, and number of rooms are four major factors to influence housing sale price.

- ***Scatter Plot***









From above four scatter plots, we can see that “Living Area” have the most obvious relationship with sale price, with the increasing living area, the sale price grows higher and higher. The “Lot Frontage” and “Lot Area” are two factors also have a little positive relationship. While due to the total number of rooms change range is not continuous, it seems to show little patterns with the sale price, in which we cannot easily make a conclusion that the number of total rooms has a direct influence on the housing sale price.

According to the analysis above, we select “Living Area” as the predictor variable in our project. In the following paragraph, we will use the statistic way to analyze how living area will affect housing sale price.

Section I

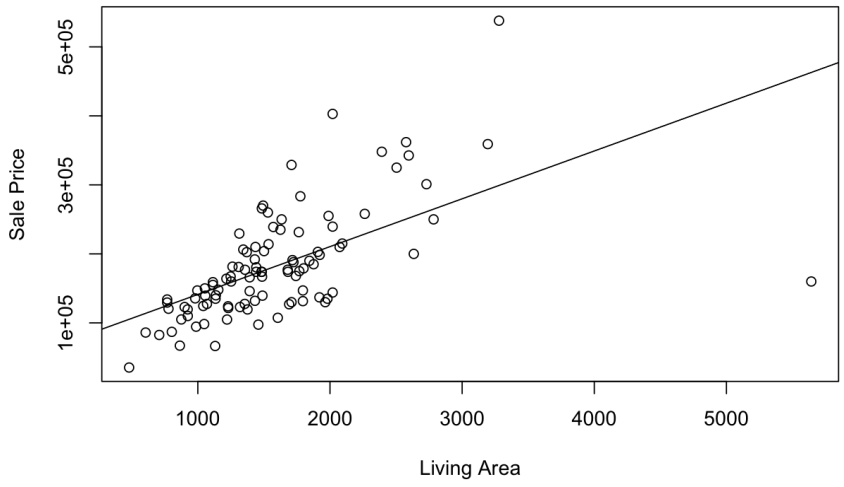
- ***Model Statement***

The regression model can reflect the relationship of two variables, which is represented as:

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

Where β_0 is the y-intercept, β_1 is the slope, ε is the random error.

In this project, our model are shown below:



The fitted model is:

$$y = 69.336x + 71901.486$$

Which means when area increase by 1 square foot, the estimate price will increase by 69.336 dollar.

- ***Model Analysis***

In this part, we calculate the correlation, ANOVA analysis and regression model summary.

```
# correlation of data
p<-cor(linear_r1, use = "all.obs", method = "pearson")

# ANOVA Table
anova(y1)
```

The correlation of this model is 0.22272377, which shows some positive relationship, but not very strong. And the following table shows ANOVA analysis, which includes three most important parameters:

MSE - 74386000000

MSE is mean squared error, tells how close a regression line is to a set of points. It measures the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value, also the distances from the points to the regression line.

SSE – 217920000000

SSE is the sum of squared estimate of errors (SSE), which is deviations predicted from actual empirical values of data.

F Value – 53.11

F value is the ratio of the mean regression sum of squares divided by the mean error sum of squares. And the value of Pr (>F) is the probability that the null hypothesis is true – that is, all coefficients are zero. In this model, the Pr (>F) is small enough to represent there must be relevance between estimate price and its area.

```
## Response: linear_r1$SalePrice
##                     Df     Sum Sq    Mean Sq F value    Pr(>F)
## linear_r1$GrLivArea  1 2.1792e+11 2.1792e+11   53.11 8.116e-11 ***
## Residuals           98 4.0210e+11 4.1031e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Regression summary shows all the coefficients and residuals analysis, which comprise of the model. And R-squared and p-value evaluate the fitness of the model. In this model, we get 0.3515 in R-squared and very small number in p-value, it shows the model is kind of reasonable but not good enough.

```
# Summary of regression model
summary(y1)
```

```
##
## Call:
## lm(formula = linear_r1$SalePrice ~ linear_r1$GrLivArea, data = linear_r1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -303094  -35599   -4848   27661  238746
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         71901.486  16316.590   4.407 2.69e-05 ***
## linear_r1$GrLivArea    69.336      9.514   7.288 8.12e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 64060 on 98 degrees of freedom
## Multiple R-squared:  0.3515, Adjusted R-squared:  0.3448
## F-statistic: 53.11 on 1 and 98 DF,  p-value: 8.116e-11
```

Generally, we can reject the null hypothesis when p-value $\leqslant 0.05$, as it indicates strong evidence against the null hypothesis, so this model is acceptable.

Section II

- ***Inferences on the True Line and Prediction***

Estimate the slope β_1 using a two-sided 100(1-α)% C.I. The upper and lower bounds are as below:

$$\underset{\text{Upper Bound}}{\left(-\infty, b_1 + t_{\alpha,n-2} s.e.\{b_1\}\right)} \qquad \underset{\text{Lower Bound}}{\left(b_1 - t_{\alpha,n-2} s.e.\{b_1\}, \infty\right)}$$

Then we need to test if the slope β_1 is significant at the α level. $H_0: \beta_1 = \beta$ vs. $H_1: \beta_1 \neq \beta$.
Using a test statistic, we reject if H_0 if

$$|t^*| = \frac{|b_1 - \beta|}{s.e.\{b_1\}} > t_{\frac{\alpha}{2},n-2}$$

Using the C.I., we reject H_0 if β is not in the two- sided 100(1-α)% C.I.

$$s.e.\{b_1\} = \sqrt{\frac{MSE}{S_{XX}}} = \sqrt{\frac{74386000000}{45328583}} = 1641.039$$

$$t_{.025,98} = 1.984$$

$$b_1 \pm t_{0.025,98} s.e.\{b_1\} = 69.336 \pm (1.984) * 1641.039$$
$$= (-3186.485, 3325.157)$$

We are 95% confident that on average sale price increases by between -3186.485 and 3325.157 for each unit increase in living area.

Then we need to test if the slope β_1 is significant at the 0.05 level. $H_0: \beta_1 = 0$ vs. $H_1: \beta_1 \neq 0$.
Using a test statistic, we reject if H_0 if

$$t^* = \frac{b_1 - 0}{s.e.\{b_1\}} = \frac{69.336 - 0}{1641.039} = 0.0422$$

Since $|0.0422| < 1.984$, and 0 is in the confidence interval (-3186.485,3325.157), we cannot reject H_0.

- ***Inferences on the True Line and Prediction***

Estimate the mean response using a two-sided 95% confidence interval when living area is 1577:

$$\hat{y}|_{x=1577} = 69.336x + 71901.486 = 69.336 * 1577 + 71901.486 = 181244.358$$

Standard errors for $\hat{y}$ when living area is 27273.8

$$s.e.\{\hat{Y}|_{x=1577}\} = \sqrt{MSE\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}\right]}$$

$$= \sqrt{74386000000\left[\frac{1}{100} + \frac{(1577-1577.3)^2}{45328583}\right]}$$

$$= 27273.8$$

Use two-sided 95% C.I.'s to estimate the mean response when living area is 1577.

$$\hat{y}|_{x=1577} \pm t_{.025,98}\, s.e.\{\hat{Y}|_{x=1577}\} = (127133.139, 235355.577) \rightarrow width = 108222.4$$

We are 95% confident that the mean house sale price for a living area of 1577 is between 127133.139 and 235355.577.

Estimate a new response using a two-sided 95% confidence interval when living area is 1577:

$$p.e.\{\hat{y}|_{x=1577}\} = \sqrt{MSE + (s.e.\{\hat{y}|_{x=1577}\})^2} = 274098.267$$

$$\hat{y}|_{x=1577} \pm t_{.025,98}\, p.e.\{\hat{Y}|_{x=1577}\} = (-362566.604, 725055.320) \rightarrow width = 1087621.9$$

We are 95% confident that a new observation of house sale price for a living area of 1577 will lie between -362566.604 and 725055.320.

Section III

- ***Model assumption***

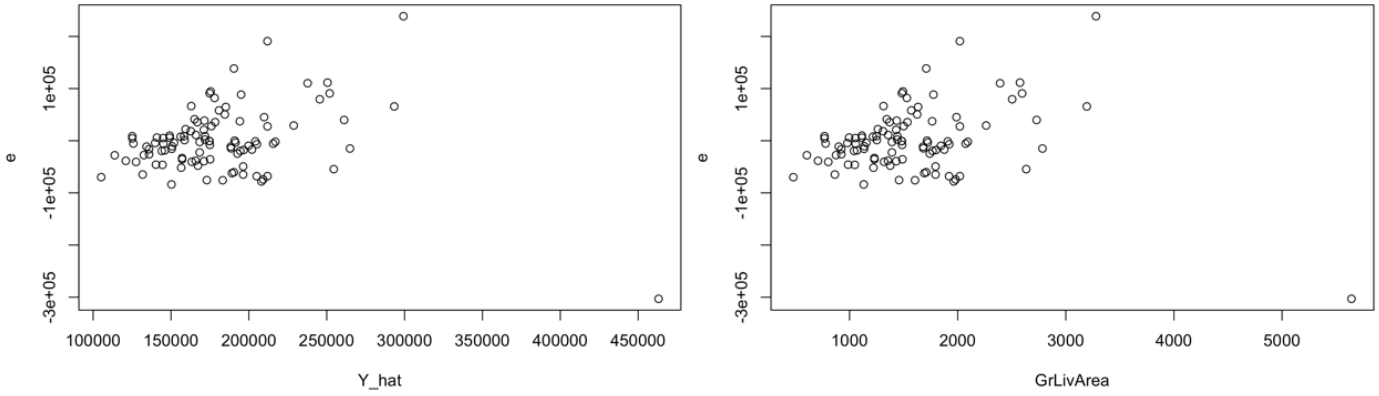
Residual: $e_i = y_i + \hat{y}_i$

The purpose of residual analysis is to verify the model assumption:

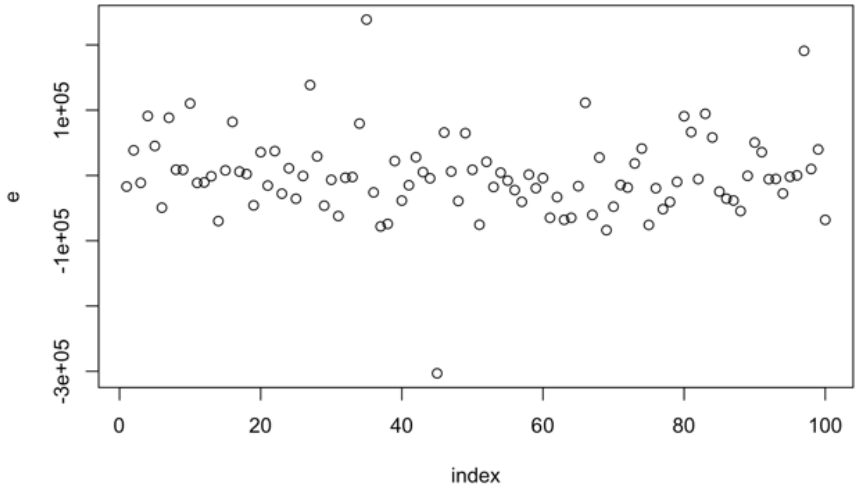
a) A linear model is reasonable
b) The residuals have constant variance
c) The residuals are normally distributed
d) The residuals are unconnected
e) No outliers

We conducted some useful plots shown as below:

1. $e_i\ VS.\ \hat{y}_i, e_i\ VS.\ x_i$



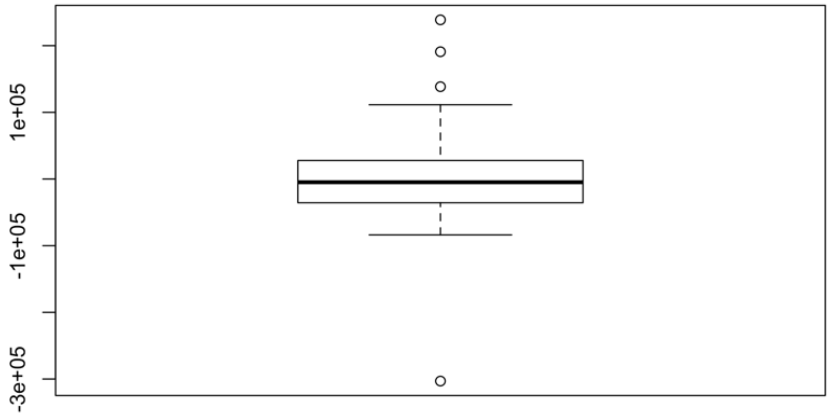
The plots above are in a funnel shape, it shows that the variance is not constant in some circumstances. Also, the plots show a x-outlier, we could eliminate it by reducing the x-range, but we decided not remove it because the x-outlier data is valid and does not interfere the normality of the residual.



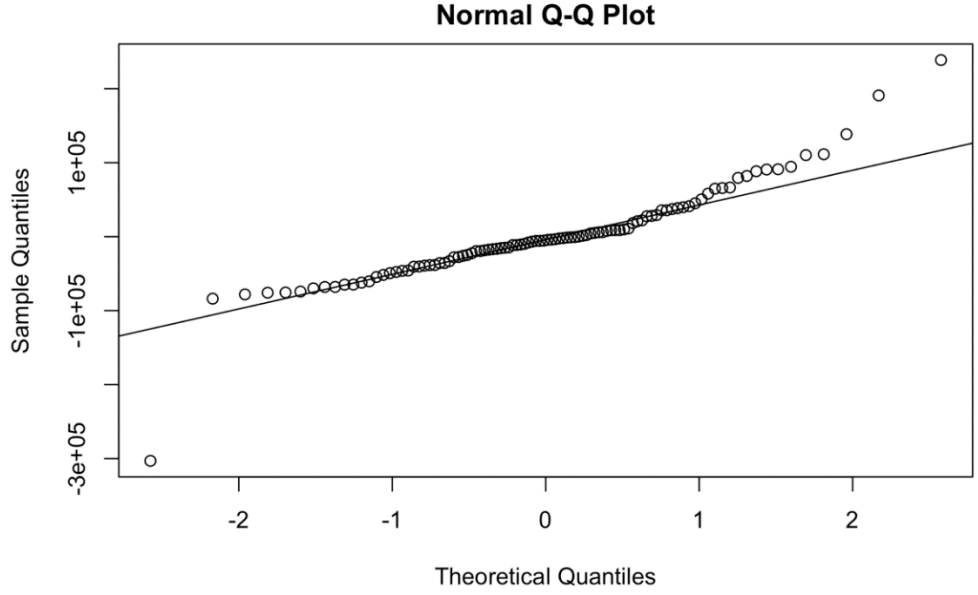
2. $e_i\ VS.\ index\ i$

The time series plot shown above is randomly distributed with no trend, which means there is no serial connection, the uncorrelated assumption is not violated. There are several y-outliers, we determined them to be valid, should not be removed.

3. e boxplot



The plot above shows the boxplot for e, it indicates that there are four outliers in the dataset, we decided not to remove any of them, because they do not violate the normality of the model.



The quantile-quantile plot (Q-Q plot) is shown above, which appears to be a relatively safe model assumption. Compared to the normal distribution, the points seem to fall into a straight line, most of the data are in the center, the graph is right skewed, meaning that most of the data is located on the left with a tail of data extending out to the right.

For transformation and remedial analysis purpose, we tried to find Correlation Between Sorted Residuals And Expected Values, also conducted a Modified-Levene Test.

```r
#combineing dataset with residuals
new_data <- cbind(linear_r1,y1$fitted.values,y1$residuals)

names(new_data)[3:4] <- c("Fitted_values","residuals")

#standard errors for residuals

std_err = summary(y1)$sigma

#finding expected value
n=nrow(new_data)
expected_value = sapply(1:n,function(k) std_err*qnorm((k-0.375)/(n+0.25)))

rho <- cor(expected_value,sort(y1$residuals))
```

Correlation - 0.9344119

We found that the Correlation Between Sorted Residuals. And Expected Values is 0.9344119, compared with the value we got from Critical Values and Coefficient of Correlation table, n =100 and Level of Significance = 0.05, the Critical Values = 0.987. The correlation we found is smaller than the Critical Values, therefore the normality assumption is not supported.

```r
#modified-levene test

Group1 <- new_data[which((new_data$GrLivArea) < median(new_data$GrLivArea)), "residuals"]
Group2 <- new_data[which((new_data$GrLivArea) >= median(new_data$GrLivArea)), "residuals"]

#calculate mean absolute deviation
M1 <- median(Group1)
M2 <- median(Group2)

#mean absolute deviation
D1 <- sum(abs(Group1-M1)/length(Group1))
D2 <- sum(abs(Group2-M2)/length(Group2))

#pooled standard Error
s_levene <- sqrt((sum((abs(Group1-M1)-D1)^2) + sum((abs(Group2-M2)-D2)^2)) /(n-2))
```

Modified-Levene Test

We conducted a Modified-Levene test to check whether the error variance is constant or not by using hypothesis testing.

H0: Error Variance is Constant

H1: Error Variance is not Constant

The result of the test is shown below:

p-value = 0.2227238 > alpha (0.05)

Thus, we cannot reject H0, the Error Variance is Constant.

```r
#caculate absolute value of the test statistics
t_levene <- abs((D1-D2)/(s_levene-sqrt((1/length(Group1))+(1/length(Group2)))))

#find p-value for the above t-value
p_levene = pt(t_levene,n-2,lower.tail = FALSE)
# p = 0.2227238
# p > alpha, cannot reject H0, error variance is constant
```

The assumption of the linear model is reasonable but not fit enough, the residuals are normally distributed and the residuals are unconnected are satisfied proofed by the plots above. However, the $e_i\ VS.\ \hat{y}_i, e_i\ VS.\ x_i$ plot show that the residuals do not have constant variance, and the e boxplot shows there are several outliers appear in the distribution. The Correlation Between Sorted Residuals and Expected Values for the dataset is smaller than Critical Values, therefore the correlation is not normally distributed. The Modified-Levene Test result support that the Error Variance is Constant.

Final discussion

In order to figure out what kind of factor will affect house price the most, we studied the relationship between sale price and living area, and finally get the equation of $y = 69.336x + 71901.486$.

With the F test, we prove that the parameters of the regression is significant because the Pr (>F) is small enough to represent there must be relevance between estimate price and its area. With the residual analysis, we verified the model assumption, the residuals are normally distributed and the residuals are well satisfied. While there are several outliers appear in the model. The Correlation Between Sorted Residuals. And Expected Values for the dataset is smaller than Critical Values, therefore the correlation is not normally distributed. And the Modified-Levene Test result also shows the Error Variance is Constant.

In conclusion, this simple linear regression proves the positive relationship between sale price and living area, and with per 1 square foot increasing, the estimate price will increase by 69.336 dollar. However, there are must many other factors will influence the estimate price, such as location, renovation year and number of rooms. It will be more accurate to conduct a multiple regression to make prediction.

Appendices I

Id	LotFrontage	LotArea	GrLivArea	Total rooms	SalePrice
1160	76	9120	1876	6.5	185000
217	65	8450	1436	6	210000
1393	68	7838	900	5	123000
1240	64	9037	1484	4	265900
1055	90	11367	1989	6.5	255000
1325	75	9986	1795	4	147000
1348	93	15306	1776	6	283463
220	43	3010	1248	4	167240
206	99	11851	1442	5	180500
991	82	9452	2392	6.5	348000
408	63	15576	1680	6	177000
675	80	9200	1136	5	140000
341	85	14191	1908	6.5	202900
917	50	9000	480	2	35311
484	32	4500	1216	5	164000
1362	124	16158	1530	6	260000
1342	66	13695	1114	5	155000
1192	24	2645	1441	4.5	174000
233	21	1680	987	3.5	94500
63	44	6442	1370	4	202500
1103	70	7000	1134	3	135000
140	65	15426	1764	6.5	231500
529	58	9098	605	4	86000
1274	124	11512	1357	4	177000
824	60	9900	1489	5	139500
910	149	12589	1484	5.5	174000
886	50	5119	1709	5	328900
404	93	12090	2263	6.5	258000
647	60	7200	1048	4	98300
363	64	7301	1922	7	198500
381	50	5000	1691	5	127000
537	57	8924	1724	5.5	188000
831	80	11900	1392	5.5	166000
567	77	11198	2504	6.5	325000
770	47	53504	3279	8.5	538000

595	88	7990	924	4	110000
1235	55	8525	1964	5.5	130000
1131	65	7804	1981	7	135000
2	80	9600	1262	5.5	181500
1324	50	5330	708	3	82500
1387	80	16692	2784	9.5	250000
144	78	10335	1501	6	204000
130	69	8973	1053	5	150000
807	75	9750	980	5	135500
1299	313	63887	5642	7.5	160000
609	78	12168	3194	6.5	359100
1086	73	9069	996	4	147000
741	60	9600	1432	4	132000
469	98	11428	1634	5	250000
953	60	7200	768	3.5	133900
659	78	17503	1458	4.5	97500
193	68	9017	1431	6	192000
380	60	8123	1800	5.5	179000
548	54	7244	768	4	129500
660	75	9937	1486	5	167000
579	34	3604	1392	4	146000
374	79	10634	1319	5	123000
1092	24	2160	1252	5.5	160000
800	60	7200	1768	5.5	175000
879	88	11782	1155	5	148000
621	45	8248	864	4	67000
957	24	1300	1229	3.5	124000
779	60	8400	2020	6	144000
580	81	12150	1795	6	131500
1177	37	6951	923	5	119500
703	82	12438	2576	6.5	361919
194	24	2522	1709	5	130000
1134	80	9828	2020	5.5	239500
875	52	5720	1131	4.5	66500
449	50	8600	1376	5	119500
136	80	10400	1682	5	174000
1422	53	4043	1069	4	127500
811	78	10140	1309	5.5	181000
421	78	7060	1344	6	206300

75	60	5790	1605	5	107400
510	80	9600	1041	5	124500
1186	60	9738	1221	6	104900
326	50	5000	803	3	87000
110	105	11751	1844	5	190000
321	111	16259	2596	7.5	342643
1218	72	8640	1314	6	229456
1457	85	13175	2073	6	210000
765	30	9549	1494	4.5	270000
1376	89	10991	1571	5	239000
992	121	17671	1742	5.5	168000
996	51	4712	1230	5.5	121600
166	62	10106	1355	7	127500
1351	91	11643	2634	10.5	200000
858	65	8125	1481	5.5	174000
252	44	4750	1625	5.5	235000
568	70	10171	1535	5	214000
634	80	9250	1056	4.5	139400
269	71	6900	778	3	120500
147	51	6120	875	4	105000
1410	46	20544	2093	5.5	215000
653	70	8750	1716	4.5	191000
516	94	12220	2020	6.5	402861
19	66	13695	1114	5.5	159000
323	86	10380	2730	7.5	301000
1276	95	11345	1920	6	137000