RA Task Qianhui Fang

Task 1. Produce a summary statistics table at event level (note that our current data is at action-level). The table should include mean, standard deviation, minimum, median, and maximum. The variables to be summarized include property age (using the variable "built_year"), number of buyers (identified by unique "buyer_id"), number of seller revisions, duration until off market (using variable "off_market_date", in days), and sales price (using the variable "de_sales_price").

	Mean	Std.Dev.	Min	Median	Max
Property Age	40.59	32.68	0.0	32.0	156
Buyer Represented by the Platform	2.01	0.11	2.0	2.0	3
Revisions	1.65	1.36	1.0	1.0	10
Duration	40.64	64.78	1.0	15.0	406
Sales Price	460673.00	242700.03	104335.3	400652.1	1561797

By computing, the observations with missing price are removed. In addition, negative property ages have been considered unreasonable.

From the output table, we can find that:

1. The average property age is around 41 years but with a large SD value which means that the spread of ages is quite large, ranging from 32 to 156 years.

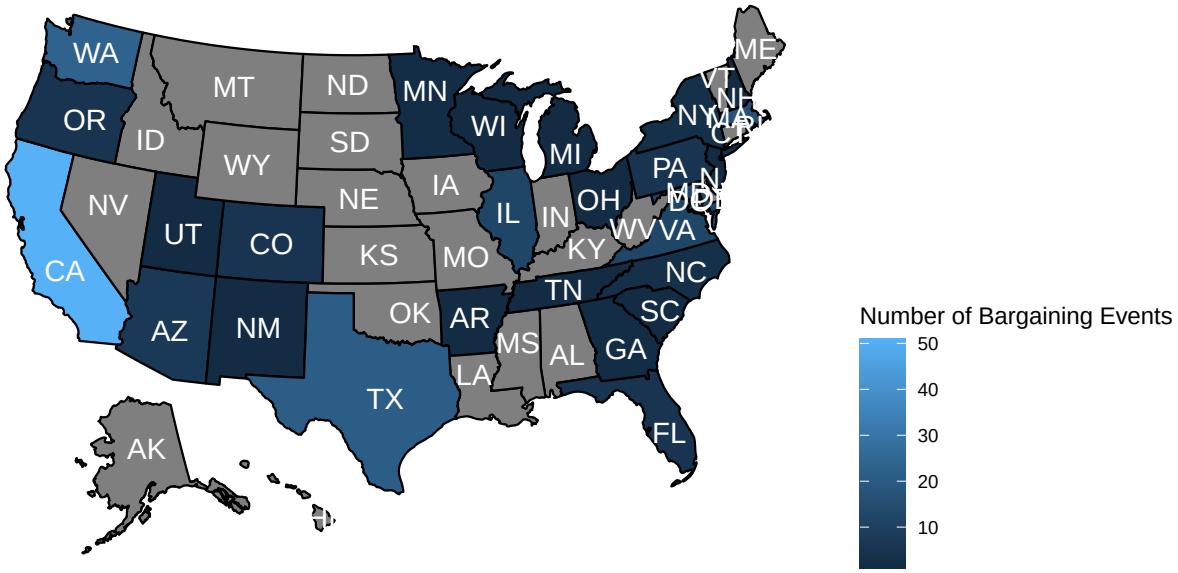
2. For most of the cases, each property has about 2 buyers.

3. More than half of the bargaining events end with between 1 to 3 rounds. There also exists an extreme case that reaches 10 rounds.

4. The mean of duration until off market is 40.64 days while the median is 15 days. The disparity shows that more than half of the properties are sold within 20 days. However, there is still many properties left unsold for a long time which leads to a higher mean.

5. The sales price ranges from 104335.3 USD to 1561797 USD with a median of 400652.1 USD.

Task 2. Draw a map with all states in the U.S. to illustrate the geographical distribution of the bargaining events in the data sample. Use colors to represent the number of bargaining events from each state in our sample. A clear legend should be provided.



From the map, we can find that:

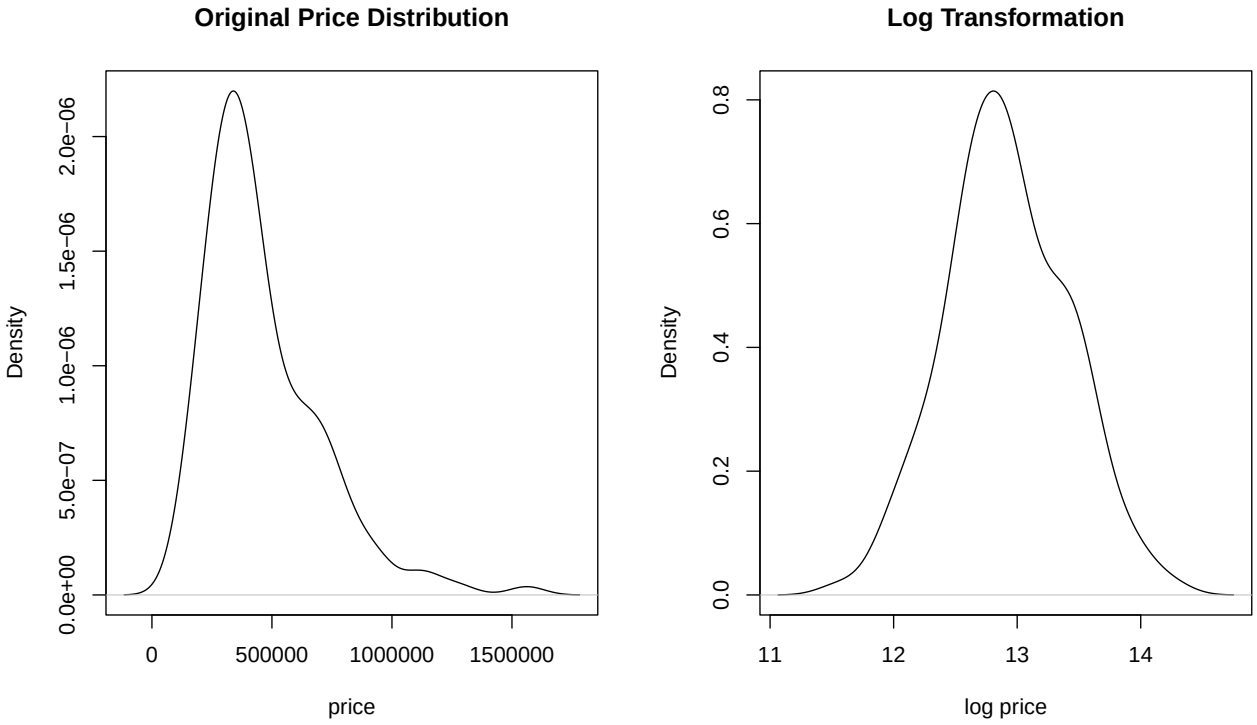
1. California has 51 bargaining events which is the highest among four states.

2. Washington state and Texas rank the second and the thrid, with 23 events and 21 events respectively.

3. Most of the states from MidWest America are relatively less involved in bargaining events.

Task 3. What factors can affect the final sales price of a property ("sales_price" in the data)?

To improve the linearity, I first logged `price` to make sure the linearity for the regressions. From the plots, we can find that the distribution of `price` after the log transformation is more normally distributed.



I selected `num_bathroom`, `num_bedrooms`, `approx_sq_ft`, `walk_score`, `bike_score`, `lot_sq_ft`,`total_num_buyers_event`, `de_original_list_price`, `age`, `revision`, `buyer_nr`, and `dur` as factors which would possibly affect the sales price.

Then, I joined everything at the event-level and replaced all the missing values in each column with the average of the rest.

To conduct variable selection, I made a Full_model and applied stepwise regression. The result left me the variables that are statistically significant. I store the result of the stepwise regression into the Stepwise_model.

Full_model: log(price) ~ age + buyer_nr + revision + dur + num_bathrooms + num_bedrooms + approx_sq_ft + walk_score + bike_score + lot_sq_ft + total_num_buyers_event + de_original_list_price

Stepwise_model: log(price) ~ de_original_list_price + dur + total_num_buyers_event + lot_sq_ft + num_bedrooms + approx_sq_ft + walk_score

Below is the summary of the Stepwise_model:

```
##
## Call:
## lm(formula = log(price) ~ de_original_list_price + dur + total_num_buyers_event +
##     lot_sq_ft + num_bedrooms + approx_sq_ft + walk_score, data = model_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.62003 -0.04631  0.03049  0.09677  0.30474
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              1.202e+01  5.513e-02 218.008  < 2e-16 ***
## de_original_list_price   1.919e-06  5.930e-08  32.355  < 2e-16 ***
## dur                     -9.912e-04  2.059e-04  -4.815 3.71e-06 ***
## total_num_buyers_event   1.223e-02  3.380e-03   3.618 0.000411 ***
## lot_sq_ft                5.194e-07  3.460e-07   1.501 0.135489
## num_bedrooms            -4.579e-02  1.701e-02  -2.692 0.007943 **
## approx_sq_ft             6.367e-05  2.223e-05   2.864 0.004808 **
## walk_score               8.519e-04  4.702e-04   1.812 0.072140 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1506 on 143 degrees of freedom
## Multiple R-squared:  0.9117, Adjusted R-squared:  0.9074
## F-statistic:   211 on 7 and 143 DF,  p-value: < 2.2e-16
```

From the summary of the Stepwise_model, we can find that `lot_sq_ft` and `walk_score` are not statistically signifianct as their p-value is greater than 0.05. Therefore, I removed them from the model and stored the rest of the variabels in to the Restricted_model.

Restricted_Model = log(price) ~ de_original_list_price + dur + total_num_buyers_event + num_bedrooms + approx_sq_ft

```
##
## Call:
## lm(formula = log(price) ~ de_original_list_price + dur + total_num_buyers_event +
##     num_bedrooms + approx_sq_ft, data = model_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.63365 -0.04630  0.03238  0.08762  0.31657
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              1.209e+01  4.473e-02 270.227  < 2e-16 ***
## de_original_list_price   1.947e-06  5.837e-08  33.355  < 2e-16 ***
## dur                     -9.747e-04  2.062e-04  -4.727 5.35e-06 ***
## total_num_buyers_event   1.374e-02  3.335e-03   4.119 6.38e-05 ***
## num_bedrooms            -5.534e-02  1.668e-02  -3.318  0.00114 **
```

```
## approx_sq_ft            5.902e-05  2.135e-05   2.764  0.00645 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1524 on 145 degrees of freedom
## Multiple R-squared:  0.9084, Adjusted R-squared:  0.9052
## F-statistic: 287.4 on 5 and 145 DF,  p-value: < 2.2e-16
```

Below is the regression output(Table 1) and my interpretation.

According to the $R^2 = 0.905$, the model works quite well. About 91% variation of price could be explained by the predictors' variation. Moreover, on average:

1. Increasing the duration until off market by 1 day leads to a 0.1% decrease in the price of the property.

2. Increasing the number of bedrooms by 1 day leads to a 5.5% decrease in the price of the property.

3. Increasing the size of the living space in the house by 1 square feet leads to a 0.01% decrease in the price of the property.

4. Increasing the number of bargaining events by 1 leads to a 1.4% decrease in the price of the property.

5. The initial listing price the seller wants does not seem to affect the price of the property by a significant amount.

6. The constant is fairly large compared to other variables. It means that we did not count in many potential factors which can affect the price.

In conclusion, from the regression output, I found that the duration until off market, the number of bedrooms, the size of the living space, the number of bargaining events, and the initial listing price would affect the final sales price of a property.

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Fri, Feb 25, 2022 - 10:14:23

Table 1: Models

	Dependent variable:	
	log(price)	
	(1)	(2)
age	0.0002	
	(0.0005)	
buyer_nr	0.055	
	(0.112)	
revision	−0.001	
	(0.012)	
dur	−0.001***	−0.001***
	(0.0003)	(0.0002)
num_bathrooms	0.007	
	(0.024)	
num_bedrooms	−0.048**	−0.055***
	(0.019)	(0.017)
approx_sq_ft	0.0001**	0.0001***
	(0.00003)	(0.00002)
walk_score	0.001	
	(0.001)	
bike_score	−0.0002	
	(0.001)	
lot_sq_ft	0.00000	
	(0.00000)	
total_num_buyers_event	0.012***	0.014***
	(0.003)	(0.003)
de_original_list_price	0.00000***	0.00000***
	(0.00000)	(0.00000)
Constant	11.906***	12.086***
	(0.237)	(0.045)
Observations	151	151
R^2	0.912	0.908
Adjusted R^2	0.904	0.905
Residual Std. Error	0.153 (df = 138)	0.152 (df = 145)
F Statistic	119.305*** (df = 12; 138)	287.441*** (df = 5; 145)
Note:		*p<0.1; **p<0.05; ***p<0.01