ISOM 670 Business Statistics, Final Quiz August 2019

The Excel spreadsheet **housedata.xls** contains data on the sales of 950 single-family homes in Springfield, MA. We wish to explain and predict the price of a single-family home (Y, in thousands of dollars) using the following predictor variables:

Data Description

Variable Name	Description	**House of interest**
s_p	Sale price in dollars	**?**
inv	Sale date inventory of homes on market	**100**
bath	Number of bathrooms	**2**
ltsz	Lot size in acres	**.25**
hssz	Sq. ft. of living area	**1200**
bsemt	1 if basement, 0 otherwise	**0**
a_c	1 if central a/c, 0 otherwise	**1**
f_place	1 if fireplace, 0 otherwise	**0**
garsz_a	1 if garage, 0 otherwise	**1**
dinsp	1 if dining space, 0 otherwise	**1**
dw	1 if dishwasher, 0 otherwise	**1**
dr	1 if dining room, 0 otherwise	**0**
fr	1 if family room, 0 otherwise	**0**
age5	1 if age <= 5 yrs, 0 otherwise	**1**
stl10	1 if 1 story house, 0 otherwise	**1**
bdrms	Number of bedrooms	**4**

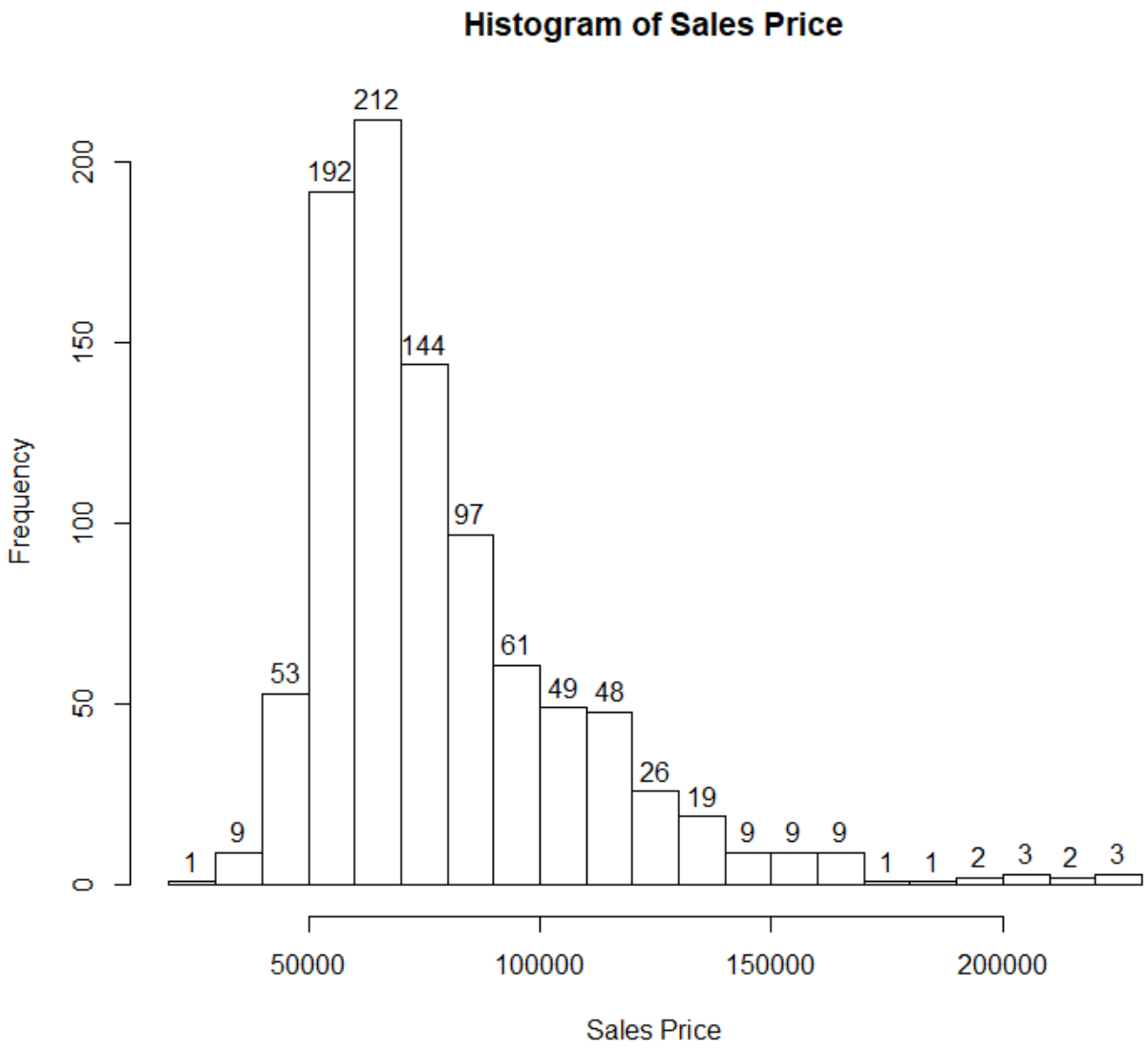
1) Calculate simple descriptive statistics for "Sales Price"

```
> describe(data$s_p)
   vars   n     mean       sd median  trimmed      mad   min    max  range skew
X1    1 950 79037.11 29169.78  70360 74918.59 20115.92 29864 222680 192816 1.71
   kurtosis     se
X1     4.05 946.39
> summary(data$s_p)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  29864   59278   70360   79037   90741  222680
```

2) and **comment.**

Sales Price has a mean of 79037 USD with a standard deviation of 29170. The median price is 70360 USD, which is lower than the mean price, indicating that the distribution of Sales Price is right-skewed. The minimum price is 29864 USD, while the maximum price is 222680 USD. The 1st quantile is 59278 USD, indicating that 25% of the whole dataset were sold below this number. The 3rd quantile is 90741 USD, far away from the max (222680 USD). This tells us that the highest 25% sales were sold with very expensive prices.

3) Construct a clear well labeled Histogram of "Sales Price"

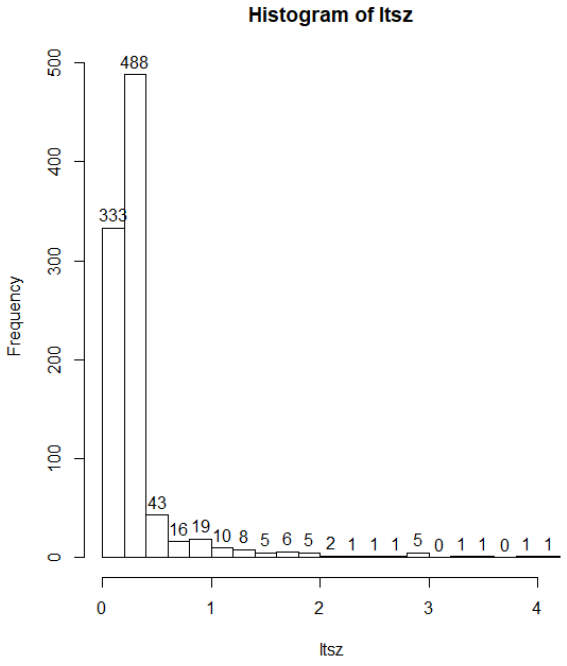

Histogram of Sales Price
Frequency
0
50
100
150
200
1
9
53
192
212
144
97
61
49
48
26
19
9
9
9
1
1
2
3
2
3
50000
100000
150000
200000
Sales Price

4) and **comment** on what you see.
The highest frequency is the range of 60000 – 70000 with 212 sales. The lowest frequency is 1. Only 10 sales are below 40000 USD. Only 8 sales are above 200000 USD. Most sales happen in the range of 50000 – 100000. The distribution of Sales Price is right-skewed, indicating there're lots of high-price sales.

5) Build a regression model to predict the selling price for a home. Explain your thinking and your analytical process concisely but clearly, using specific excerpts from your data analysis where appropriate Be sure to discuss any additional steps you would like to perform if you had more time for your analysis (and why those steps would be important.

Data Cleansing: After browsing the dataset, I find some apparent mistakes. In observation 675, bath = 20; in observation 950, garsz_a = 2, which should be either 0 or 1; in observation 46, bdrms = -4, which should be a positive number. Since there are only 3 such erroneous data, I decide to exclude them from the dataset, reducing the available observations from 950 to 947.

Variable Selection: (1) Log(s_p) is used instead of s_p. This is because s_p is highly right-skewed and is affected




```
Call:
lm(formula = log(s_p) ~ inv + factor(bath) + log(ltsz) + hssz +
    bsemt + a_c + f_place + garsz_a + dw + dr + fr + age5 + stl10 +
    factor(bdrms), data = partdata)

Residuals:
     Min       1Q   Median       3Q      Max
-0.90336 -0.11623 -0.00758  0.10970  0.73809

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.065e+01  1.070e-01  99.490  < 2e-16 ***
inv            -3.222e-04  1.163e-04  -2.770 0.005725 **
factor(bath)1.5 2.381e-02  2.205e-02   1.080 0.280481
factor(bath)2   6.897e-02  2.107e-02   3.272 0.001106 **
factor(bath)2.5 2.202e-01  2.703e-02   8.146 1.21e-15 ***
factor(bath)3   1.570e-01  3.061e-02   5.131 3.52e-07 ***
factor(bath)3.5 4.880e-01  5.229e-02   9.333  < 2e-16 ***
factor(bath)4   3.540e-01  8.101e-02   4.370 1.38e-05 ***
log(ltsz)       1.435e-01  1.058e-02  13.570  < 2e-16 ***
hssz            1.188e-04  2.267e-05   5.240 1.99e-07 ***
bsemt           1.404e-01  2.505e-02   5.607 2.71e-08 ***
a_c             3.417e-02  1.756e-02   1.945 0.052052 .
f_place         1.126e-01  1.463e-02   7.699 3.52e-14 ***
garsz_a         4.826e-02  3.179e-02   1.518 0.129384
dw              2.638e-02  1.866e-02   1.414 0.157679
dr              8.315e-02  1.547e-02   5.375 9.71e-08 ***
fr             -2.077e-03  1.753e-02  -0.118 0.905725
age5            1.119e-01  1.730e-02   6.464 1.65e-10 ***
stl10          -1.495e-02  1.315e-02  -1.138 0.255602
factor(bdrms)2  1.195e-01  9.410e-02   1.270 0.204371
factor(bdrms)3  2.188e-01  9.258e-02   2.363 0.018311 *
factor(bdrms)4  3.090e-01  9.404e-02   3.286 0.001053 **
factor(bdrms)5  3.371e-01  9.729e-02   3.465 0.000555 ***
factor(bdrms)6  4.407e-01  1.634e-01   2.697 0.007117 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1803 on 923 degrees of freedom
Multiple R-squared:  0.7001,    Adjusted R-squared:  0.6926
F-statistic: 93.67 on 23 and 923 DF,  p-value: < 2.2e-16
```

by a few large outliers. Log(s_p) can reduce the impact of outliers. Log(ltsz) is used because of the same logic. (2) Factor(bath) and factor(bdrms) are used instead. (3) Stl10, fr, dw, and garsz_a are dropped due to statistical insignificance.

Model: lm(formula = log(s_p) ~ inv + factor(bath) + log(ltsz) + hssz + bsemt + a_c + f_place + dr + age5 + factor(bdrms), data = partdata)

```
Call:
lm(formula = log(s_p) ~ inv + factor(bath) + log(ltsz) + hssz +
    bsemt + a_c + f_place + dr + age5 + factor(bdrms), data = partdata)

Residuals:
     Min       1Q   Median       3Q      Max
-0.89750 -0.11122 -0.00978  0.11157  0.74896

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.068e+01  1.016e-01 105.137  < 2e-16 ***
inv             -2.862e-04  1.145e-04  -2.499 0.012624 *
factor(bath)1.5  3.174e-02  2.102e-02   1.510 0.131376
factor(bath)2    7.673e-02  2.015e-02   3.808 0.000149 ***
factor(bath)2.5  2.294e-01  2.592e-02   8.850  < 2e-16 ***
factor(bath)3    1.637e-01  2.963e-02   5.523 4.32e-08 ***
factor(bath)3.5  4.982e-01  5.167e-02   9.643  < 2e-16 ***
factor(bath)4    3.663e-01  8.058e-02   4.546 6.18e-06 ***
log(ltsz)        1.448e-01  1.055e-02  13.726  < 2e-16 ***
hssz             1.169e-04  2.206e-05   5.301 1.44e-07 ***
bsemt            1.429e-01  2.479e-02   5.765 1.11e-08 ***
a_c              4.564e-02  1.672e-02   2.729 0.006476 **
f_place          1.175e-01  1.429e-02   8.221 6.76e-16 ***
dr               8.947e-02  1.507e-02   5.939 4.06e-09 ***
age5             1.156e-01  1.710e-02   6.761 2.42e-11 ***
factor(bdrms)2   1.164e-01  9.391e-02   1.239 0.215538
factor(bdrms)3   2.189e-01  9.205e-02   2.378 0.017619 *
factor(bdrms)4   3.123e-01  9.326e-02   3.349 0.000844 ***
factor(bdrms)5   3.402e-01  9.669e-02   3.519 0.000455 ***
factor(bdrms)6   4.527e-01  1.632e-01   2.774 0.005652 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1806 on 927 degrees of freedom
Multiple R-squared:  0.698,     Adjusted R-squared:  0.6918
F-statistic: 112.8 on 19 and 927 DF,  p-value: < 2.2e-16
```

Additional Steps: (1) Houses with multiple stories are harder and more expensive to build than one-story houses. Thus, it's probably helpful to treat stl10=0 and stl10=1 as two different groups. Also, people buying multiple-story houses may want different things than those buying one-story houses. It might be a good idea to develop two different models for these two groups. (2) Gathering and combining more data can be useful. For example, if I know the exact numbers of stories for the stl10=0 group, I should be able to do factor(stl10). Since the inv is different, the dataset is likely to be gathered in a long period of time. Thus, the CPI data or other economic data can explain the difference in s_p caused by external factors.

6) What is your BEST -MOST COMPLETE answer to what the house of interest listed above will cost?

```
> predict.lm(modela, newdata=data.frame(inv=100,bath=2,ltsz=0.25,hssz=1200,bsemt=0,a_c=1,f_place=0,garsz_a=1,dw=1,dr=0,
fr=0,age5=1,stl10=1,bdrms=4),interval="prediction",level=0.95)
       fit      lwr      upr
1 11.13872 10.77757 11.49987
```

Previous results are in logarithm. Thus, the actual predicted s_p is exp(11.13872) = 68784 USD.

Lower-limit is exp(10.77757) = 47934 USD; higher-limit is exp(11.49987) = 98703 USD.