Dr. Weihong Grace Guo
wg152@soe.rutgers.edu

Project: Housing Prices Analysis

Due on Thursday, December 10, at 11:59 PM

Project may be done **individually or in a team of 2 students**. When the project is done as a team, the group writeup must be submitted as a single copy bearing the names of both group members (who generally share the same grade unless there are extenuating circumstances).

Submit your project report and programs to Canvas.

- Report: Present your answers in a Statistical Analysis Report. The report should be clearly written with logical organization. You may include a short supplementary appendix as long as the materials in appendix are well motivated by the main body of the report.
- Program: Your programs should contain the data table(s) and your analyses and results. Acceptable formats include but are not limited to *.jmp, *.jrp, *.jrn, *.xlsx, *.xlsm, *.R, *.Rmd, *.m, *.py. If JMP is used, save your analyses by **Save Script > To Data Table**.
- Software: You may use any software available to you.

Data Set Information

The `KCHousingPrices` data set contains information on a random sample of home sales in the King County area, Washington State, between May 2014 and May 2015. The data set contains 212 rows. Each row represents a home sold. There are 15 variables in the data set:

- `id`: Unique ID for each home sold
- `price`: Selling price of the property ($1,000)
- `bedrooms`: Number of bedrooms in the house
- `bathrooms`: Number of bathrooms in the house
- `sqft_living`: Square footage of the interior living space
- `sqft_above`: The square footage of the interior housing space that is above ground level
- `floors`: Number of floors
- `view`: An index from 0 to 4 of how good the view of the property was
- `condition`: An index from 1 to 5 on the condition of the house, 1=Poor, 5=Very Good.
- `grade`: An index from 1 to 13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have a high quality level of construction and design. For more descriptions on `condition` and `grade`, see here.
- `yr_built`: The year the house was initially built
- `yr_renovated`: The year of the house's last renovation
- `latitude`: The latitude of the house
- `longitude`: The longitude of the house
- `sqft_living15`: The square footage of interior living space for the nearest 15 neighbors

Data source: Raw data from Kaggle House Sales in King County, USA. I preprocessed the data by removing waterfront houses and taking a random sample of size 212 from the data.

Dr. Weihong Grace Guo
wg152@soe.rutgers.edu

Statistical Analysis

1. Data Exploration and Visualization

(a) Which of the variables are quantitative, and which are qualitative?

(b) The `price` is a key variable of interest. Provide useful summary statistics and graphical displays for `price`. Find out its distribution.

(c) Examine all the other variables (except `id`). Produce some numerical and graphical summaries of the data. Describe the shapes of the distributions.

- Note: For discrete variables, provide bar charts and frequency distributions. For continuous variables, provide histograms and summary statistics to characterize the shape, center, and spread of the distributions.

(d) Produce a scatterplot matrix which includes all of the variables (except `id`) in the data set. You may also produce scatterplots for just a subset of the variables. Create some plots highlighting the relationships among the variables. Describe your findings.

(e) Suppose that we wish to predict `price` on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting `price`? Justify your answer.

2. Statistical Inference

(a) Find the 95% confidence interval for the average selling price of King County houses, and the 95% confidence interval for the average square footage of King County houses. What can you conclude from these intervals?

(b) It is claimed that houses with basements have a higher average selling price than houses without basements. Test this claim using $\alpha = 0.05$. Here are the steps:

i. Create a binary variable, `has basement?`, that contains a "Y" if the house has basement (`sqft_living` is larger than `sqft_above`), and an "N" otherwise.

ii. Test the hypothesis that $\mu_{\text{basement}} = \mu_{\text{no basement}}$ versus the alternative that $\mu_{\text{basement}} > \mu_{\text{no basement}}$. What is the P-value for this test? What can you conclude from this test?

(c) Construct a two-sided confidence interval for $\mu_{\text{basement}} - \mu_{\text{no basement}}$. What can you conclude from the CI?

(d) Produce a side-by-side boxplot to compare the prices of houses with basements versus houses without basements. Comment on the plot.

3. Simple Linear Regression

(a) Perform a simple linear regression with `price` as the response and `sqft_living` as the predictor. Plot the response and the predictor. Display the least squares regression line. Analyze the results. Comment on the output. For example:

i. Is there a relationship between the predictor and the response?

ii. How strong is the relationship between the predictor and the response?

iii. Is the relationship between the predictor and the response positive or negative?

iv. How much does the house price increase for each additional square foot?

Dr. Weihong Grace Guo
wg152@soe.rutgers.edu

v. What is the predicted selling price for a 2,000 square foot home? What are the associated 95% confidence and prediction intervals?

(b) Produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit.

(c) Perform a quadratic fit to the data:

$$\text{price} = \beta_0 + \beta_1 \times \text{sqft\_living} + \beta_2 \times \text{sqft\_living}^2 + \epsilon$$

and a linear fit with logarithm transformation on `price` as the response:

$$\ln(\text{price}) = \beta_0 + \beta_1 \times \text{sqft\_living} + \epsilon.$$

Use R^2, R^2_{adj}, and diagnostic plots to compare the quadratic fit, the linear fit with log transformation, and the fit obtained in (a). Which model do you recommend?

(d) For each predictor (except `id`), fit a simple linear regression model to predict `price`. Describe your results. In which of the models is there a statistically significant association between the predictor and the response? Create some plots to back up your assertions.

4. Multiple Linear Regression

(a) Perform a multiple linear regression with `price` as the response and all other variables except `id` as the predictors. Analyze the results. Comment on the output. For instance:

i. Is there a relationship between the predictors and the response?

ii. Which predictors appear to have a statistically significant relationship to the response?

iii. What does the coefficient for the "`latitude`" variable suggest?

iv. What does the coefficient for the "`grade`" variable suggest?

(b) Produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers?

(c) How do your results from 4(a) compare to your results from 3(d)? Create a plot displaying the univariate regression coefficients from 3(d) on the x-axis, and the multiple regression coefficients from 4(a) on the y-axis. That is, each predictor is displayed as a single point in the plot. Its coefficient in a simple linear regression model is shown on the x-axis, and its coefficient estimate in the multiple linear regression model is shown on the y-axis.

(d) Fit linear regression models with interaction effects (like `latitude`*`grade`). Do any interactions appear to be statistically significant?

(e) Try a few different transformations of the variables, such as $\log(X)$, $\sqrt{X}$, X^2. Comment on your findings.

(f) The ultimate goal is to develop the simplest model that does the best job of predicting housing prices. Consider removing non-significant predictors from the full model. Consider adding interaction terms and transformations. Write out your final model. Report R^2, R^2_{adj}, and RMSE (root mean square error).

Dr. Weihong Grace Guo
wg152@soe.rutgers.edu

5. **Classification**

Asbestos is a natural mineral composed of thin fibers. When residential construction products made with asbestos are damaged, those fibers become airborne and could pose a danger to anyone who inhales the toxic dust. Asbestos use has declined significantly since the late 1970s, when the U.S. banned spray-on asbestos and several other uses. However, many older homes still contain asbestos. Develop a classification model to predict whether or not a given house was built before 1980.

(a) Create a binary variable, `before1980`, that contains a 1 if `yr_built` is before 1980, and a 0 if the house was built in 1980 or after.

(b) Explore the data graphically in order to investigate the association between `before1980` and the other variables. Which of the other variables seem most likely to be useful in predicting `before1980`? Scatterplots and boxplots may be useful tools to answer this question. Describe your findings.

(c) Split the data into a training set and a test set.

(d) Perform logistic regression on the training data in order to predict `before1980` using the variables that seemed most associated with `before1980` in (b). What is the test error of the model obtained?

(e) Perform K-Nearest Neighbors on the training data, with several values of K, in order to predict `before1980`. Use only the variables that seemed most associated with `before1980` in (b). What test errors do you obtain? Which value of K seems to perform the best on this data set?

(f) Compare the classification results in (d) and (e). Which classification model would you recommend to predict whether or not a given house was built before 1980?

6. **Extra Credit: Creative Exercise**

Find a question that has not been investigated in the above analyses. Perform your analysis and describe your findings. This will earn you up to two extra points of course grade.

In order to earn extra points, the question of your choice must not be trivial (e.g., what is the percentage of 3-bedroom houses in the data set?). Here are just some ideas of nontrivial questions (and possible solutions):

- Perform discriminant analysis (or any other classification method) to predict `before1980`. Compare the classification results to 5(f). Which model would you recommend?
- Perform cluster analysis to identify which houses are similar to each other.
- Test the claim that houses that had renovations have a higher average selling price than houses without renovations.
- Can we predict the square footage of the home based on other variables?
- Are there possible outliers in the data? Identify outliers and remove or correct them, and then adjust the pricing prediction model accordingly.
- Download the full data set from Kaggle. It contains 21,613 homes sold between May 2014 and May 2015. Does your pricing prediction model still work?

Keep in mind that no matter how imaginative the question of your choice is, the statistical analysis should be as rigorous as possible. Do not jump to conclusions. Have fun!