


Semester Project: AD699

In many ways, this assignment is **intentionally open-ended** -- there are several parts that ask you to decide what to do, and there is not necessarily a single right or wrong answer to the question. As a group, you will decide things like which variables to use for your models, and which to ignore. The outcome matters, but you should focus more on the *process*.

"There is not necessarily a single right or wrong answer" is not the same as saying "all answers are equally wonderful." A rubric for this assignment will be posted on our class Blackboard page.

You may get results that are not "nice" or "neat" -- remember, this is real-world data, and not based on a canned textbook example. **If your model does not predict the outcome variable well, that doesn't necessarily mean that you did something wrong, or that the assignment prompt is wrong**; on the contrary, it means that you explored something, you found results, and you're reporting the results. Sometimes, it doesn't get more 'real world' than that!

The neighborhoods available for this project are:

Pudong
Huangpu District
Xuhui District
Chongming District
Jing'an District
Minhang District

You will form a team with 3-5 other classmates. Each team will choose one neighborhood group. Once you have chosen your neighborhood, e-mail your Professor to confirm the selection (each team will have its own unique neighborhood (*neighbourhood*, as spelled in the dataset), so the choices will be confirmed on a 'first come, first served' basis). These neighborhoods can be found in the dataset in the variable *neighbourhood_cleansed*.

The AD699 Teaching Team will be glad to **clarify** any questions you have about this prompt or about the rubric, which can also be found in the "Semester Project" folder. However, **the assignment itself belongs to you and your team**.

If you have a question along the lines of "Why doesn't this code work?" the teaching team will assume that you have already run this question by all of your other team members.

As you proceed, just remember this quote from Hadley Wickham, the author of the tidyverse: "Every good data analysis starts as a bad data analysis." *So get started, do it badly, and then do it better*!

Start by downloading the dataset from our class Blackboard page. It can be found in the "semester project" folder. The file name is *shanghai.csv*.

Step I: Data Preparation & Exploration (20 points)

Read your data into your local environment, and subset/filter the data so that you are dealing only with the records that pertain to your team's *neighborhood_cleansed*.

Please note: *You may wish to use read_csv() from readr, rather than read.csv(), for bringing this dataset into your environment. read_csv() uses a character encoding system that will enable all the original characters to render properly on your screen.*

I. **Missing Values**

A. Does your data contain any missing values and/or blank cells? If so, what can you do about this? Show the R code that you used to handle your missing values.

B. Write one paragraph describing what you did, and why you did it. (*Note: You may wish to deal with missing values differently for different tasks. You are not 'locked in' to a decision regarding missing values*).

II. **Summary Statistics**

A. Take a peek at your data, and then brainstorm a bit about some questions that you'd like to answer with summary statistics. To answer these questions choose any five of the summary statistics functions shown in the textbook, class slides, or anywhere else to learn a little bit about your data set.

B. Show screenshots of the results. Describe your findings in 1-2 paragraphs.

III. **Data Visualization**

A. Using ggplot, create any five plots that help to describe your data. Use five unique types of plots to do this. As you do, remember to think about the *types* of variables that you are representing with a particular plot. Think of these plots as expository (not exploratory) so be sure to include clear axis labels and plot titles.

B. Write a two-paragraph description that explains the choices that you made, and what the resulting plots show.

IV. Mapping

A. Generate a map of your neighborhood using any R mapping tool. Do any key features here seem to stand out? What are a few of the things your map shows you about the neighborhood?
B. Create a heatmap that shows the density of property listings in your district. Does this heatmap reveal any unique insights?

V. Wordcloud

A. Using the neighborhood overview column in your dataset, generate a wordcloud. What are some terms that seem to be emphasized here?

Step II: Prediction (20 points)

I. Create a multiple regression model with the outcome variable *price*.

A. Describe your process. How did you wind up including the independent variables that you kept, and discarding the ones that you didn't keep? In a narrative of at least two paragraphs, discuss your process and your reasoning. In the write-up, be sure to talk about how you evaluated the quality of your model.

B. Show a screenshot of your regression summary, and explain the regression equation that it generated.

C. Analyze any other metrics that are relevant for linear regression models. Based on these, what can you say about your model's performance in 1-2 paragraphs?

Step III: Classification (40 points)

Part I. Using **k-nearest neighbors**, predict whether a rental in your neighborhood will have some particular amenity, or combination of amenities. Use any set of numerical predictors in order to build this model. You can decide which amenity, or set of amenities, to use as your outcome variable.

(Hint: the grepl() function is worth exploring in order to perform this step).

A. Show the code you used to run your model, and the code you used to assess your model.

B. Write a two-paragraph narrative that describes how you did this. In your narrative, be sure to describe your predictor choices, and mention how you arrived at the particular *k* value that you used.

Classification, Part II. Naive Bayes

A. Using any set of predictors, build a model using the naive Bayes algorithm, with the purpose of predicting whether a particular rental will be instantly bookable. (*instant_bookable* is a logical variable in this dataset).
B. Describe a fictional apartment, and use your model to predict which bin it will fall into.
C. Show a screenshot of the code you used to build your model, the code you used to run the algorithm, and code you used to assess the algorithm.
D. Write a two-paragraph narrative that describes how you did this. In your narrative, be sure to talk about things like feature selection and testing against your training data.

Classification, Part III. Classification Tree

A. Build a classification tree that predicts the review score that a rental in your neighborhood will have. Before you can do this, you will need to first bin the review_scores_rating variable -- the number of bins you create is up to you. Do not use any of the other review_scores variables as inputs.

B. Determine the ideal size of your tree using cross-validation.

C. Using rpart.plot and your choice of graphical parameters, show your tree model here.

D. In a 1-2 paragraph write-up, describe your process. Talk about some of the features that you considered using, and your reasons why. Mention anything that you found interesting as you explored various possible models, and the process you used to arrive at the model you finished with. Talk about the relative sizes of each bin (using the number of records per bin) and how that may have impacted your model.

Step IV: Clustering (15 points)

I. Perform either a k-means analysis or a hierarchical clustering analysis in order to place rental units within your neighborhood into clusters (each observation in your dataframe is one rental unit).

****** Of any section of the project, this one offers the most opportunity to **be creative and take risks**. Think about feature engineering, too -- how/when/where can you create new variables based on existing ones?

II. Show your code and results. Name and describe each of your clusters. In 1-2 paragraphs, describe the process that you used for variable selection and model building.

III. Include at least three simple visualizations that describe your clustering model. A simple visualization can be a scatterplot, histogram, barplot, boxplot, violin plot, etc. Write 1-2 sentences for each visualization to explain what it shows.

Step V: Conclusions (5 points)

I. Write a 3-5 paragraph summary that describes your overall process and experience with this assignment. How could these findings be useful? Who could benefit from the data mining that you have performed here? You already summarized your specific steps in some other parts of the write-up, so focus on the big picture here (do not not use the conclusion to simply describe everything you did in the other parts).

Submit your final report as a PDF to Blackboard before the deadline listed on the assignment.

Also: Upload the R Script(s) that you used to generate your results.

Step VI: Presentation

I. Using a slide presentation (you may use as many slides as you wish to), summarize your findings and the overall process of analyzing this data set.

II. Upload your slides to Blackboard before your team's presentation timeslot.
    A. On the last day of the course, your team will deliver your findings in class, with a 15-minute presentation.