Exploratory Data Analysis and Visualisation/G 11374/11517
S1 2022
Summary and FAQ

This report provides a summary of the contents covered in the unit so far and provides some notes for the final project.

Unit summary

During the semester we covered the main topics summarised as follows:

- Data Preparation
- Univariate, Bivariate and Multivariate EDA and Visualisation
- Cross Sectional, Time Series and Panel Data EDA
- Additional Data Preparation and Visualisation including Outlier Detection, Clustering, Variable Creation and Feature Engineering
- Linear Modelling

These topics are introduced in a way they are linked by the observations or variables or data types. For example, for the Data Preparation topic, Missing Values, Outlier Detection and Cluster Analysis are focused on the observations i.e. rows of data frames. Univariate, Bivariate and Multivariate EDA and Visualisation are based on one, two or more than two variables i.e. columns of data frames. Cross Sectional, Time Series and Panel Data EDA are based on different data types. Variable Creation and Feature Engineering are involved not only in time series but also other data analyses. Linear Modelling is discussed together with Outlier Detection and Panel Data EDA although it is of course more widely used in analysis for two or more variables for Cross Sectional, Time Series and Panel Data.

Some plots can be grouped as listed below.

- Distribution plots: Histogram, Boxplot, Violin plot, Raincloud plot, QQ plot
- Category based comparison plots: Bar chart, Clustered bar chart
- Correlation and regression plots: Scatter plot, Rectangular bin, Heatmap, Residual plot
- Time related comparison plots: Line plot, Spaghetti plot
- Cluster plots: Dendrogram

As always, our interpretation is the key. For example, to understand and interpret a distribution plot e.g. histogram, we may look at the location, spread, shape and possible outliers. When we examine a correlation plot like a scatter plot we look at the direction, strength, pattern (linear or nonlinear) and possible outliers. For a time series line plot, we explore for trend, cyclic, seasonal and other features. For a spaghetti plot, we check for shared trends or mixed patterns reflected by the "spaghetti lines". When we deal with a dendrogram we investigate possible similar or even common features shared by the individuals in the same cluster, and dissimilar characteristics between the individuals in the different clusters.

R libraries and codes

A few packages e.g. dplyr and ggplot2 are used, so you may list the following in your R file in case they can be handy to help your EDA:

```
library(cluster)
library(corrplot)
library(devtools)
library(DMwR2)
library(dplyr)
library(egg)
library(GGally)
library(ggfortify)
library(ggplot2)
library(knitr)
library(lme4)
library(markdown)
library(modelr)
library(outliers)
library(reshape2)
library(rmarkdown)
library(shiny)
library(tibble)
library(tidyverse)
```

Selected codes can be collected as well:

```
read.csv, as.data.frame, as.factor, select, par etc
mean, var, cor etc
ggplot together with geom_boxplot, geom_violin, geom_point and geom_smooth etc
ggpairs, lm, summary, rmse etc
hclust and kmeans
```

The other R functions can be found from our lecture notes and lab solution files, and of course from the books and blogs available.

Final project FAQ

Some general questions and their answers are as follows.

Q1: There are two datasets available (train and test). When should we use which?

Answer: Please use the train dataset to complete all steps up to the end of the EDA and final preprocessing stages. This means all your data cleaning, missing value treatment, EDA and visualisations, and model fitting etc. The test dataset should only be used when you compare the actual values of SalePrice against your predicted values given by your trained models when fitted on the train dataset.

Q2: Is a neighborhood comparison important to consider or not?

Answer: This is up to you as EDA is open minded. There are 25 neighbourhoods. So, just interpret the visualisation and try to find insight, or determine whether neighbourhood is a significant variable. As you can see, one particular neighbourhood has a lot of observations, and some have very little. So try to note these and keep them in mind (especially if you find the medians of neighbourhood price. Some neighbourhoods with a low number of observations may have a very low or high median). Also, if your problems of interest involve neighbourhood, such as "which neighbourhood showed the highest growth over time" then neighbourhood would most definitely be a variable to compare and understand

Q3: Is the residual standard error (RSE) in the linear model summary the same as the metric root mean squared error (RMSE) or not?

Answer: The Residual standard error in the linear model summary given by R is the square-root of the Mean Square Error (MSE), which is the sum of squared residuals divided by n-k with n being the number of values in the test dataset and k being the number of parameters in the linear model. The Root Mean Square Error (RMSE) uses n instead of n-k by practitioners in their own code. We note that the difference between the two is very small numerically when n is very large but k is small.

Remember, with all EDA and modelling, try to find the variables and model(s) which may help you predict house prices.

Further readings

We wrap up by listing the following online texts and blogs which certainly provide useful information for the topics relevant to our Exploratory Data Analysis and Visulisation:

(1) ggplot2: Elegant Graphics for Data Analysis, by Hadley Wickham
https://ggplot2-book.org/

(2) R for Data Science, by Garrett Grolemund and Hadley Wickham
https://r4ds.had.co.nz/data-import.html

(3) R Tutorial
https://www.statmethods.net/r-tutorial/index.html

(4) Data Visualization in R
https://data-flair.training/blogs/data-visualization-in-r/

(5) Tidy Rain cloud plot, by Paula Andrea Martinez
https://orchid00.github.io/tidy_raincloudplot

(6) Raincloud plots: a multi-platform tool for robust data visualization, by M. Allen et al.
https://wellcomeopenresearch.org/articles/4-63

(7) ggplot2 scatter plots
http://www.sthda.com/english/wiki/ggplot2-scatter-plots-quick-start-guide-r-software-and-data-visualization

(8) The R Graph Gallery
https://www.r-graph-gallery.com/