EDA Checklists

EDA Checklists

- Formulate your question
- Check your data
- Automate your project workflow

Example Dataset - Boston

Boston dataset, 506 rows and 14 columns. The medv variable is the target variable, for more info and download check the following link:

https://www.kaggle.com/c/boston-housing

- What is the question you are really interested in, narrow it down to be as specific as possible:

  Build a model to predict medv.

  - Did you specify the type of data analytic question (e.g. exploration, association causality) before touching the data? Predictive modelling
  - Did you define the metric for success before beginning? mean squared error (MSE) because medv is numeric.
  - Did you understand the context for the question and the scientific or business application? Determining prices and the effect of clean air.
  - Did you record the experimental design? Only if you are collecting the data yourself, every step should be specified.
  - Did you consider whether the question could be answered with the available data? We have several variables, we have enough observations.

Checking your data - Checklist

1. Read in your data:

```
> boston <- read.csv("boston.csv")
```

2. Run str() or glimpse()

```
> str(boston)
```

```
'data.frame': 506 obs. of  13 variables:
$ crim   : num  0.00632 0.02731 0.02729 ...
$ zn     : num  18 0 0 0 0 0 12.5 12.5 ...
$ indus  : num  2.31 7.07 7.07 2.18 2.18 ...
$ chas   : int  0 0 0 0 0 0 0 0 0 0 ...
$ nox    : num  0.538 0.469 0.469 0.458 ...
$ rm     : num  6.58 6.42 7.18 7 7.15 ...
$ age    : num  65.2 78.9 61.1 45.8 54.2 ...
$ dis    : num  4.09 4.97 4.97 6.06 6.06 ...
$ rad    : int  1 2 2 3 3 3 5 5 5 5 ...
$ tax    : int  296 242 242 222 222 222 ...
$ ptratio: num  15.3 17.8 17.8 18.7 18.7 ...
$ lstat  : num  4.98 9.14 4.03 2.94 5.33 ...
$ medv   : num  24 21.6 34.7 33.4 36.2 ...
```

3. Look at the top and the bottom of your data, does that make sense? Any rows appearing as NAs by mistake?

```
> head(boston, 3)
> tail(boston, 3)
```

4. Did you check the total number of observations? "n"

```
> dim(boston)
```

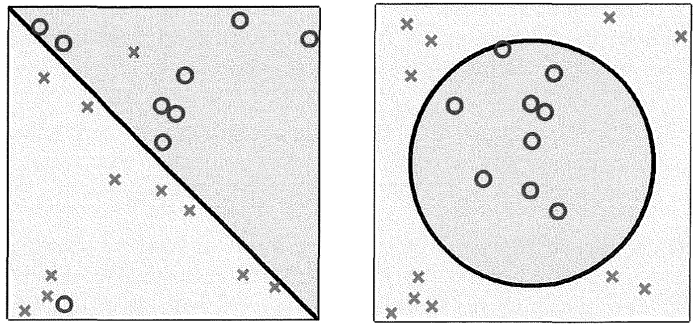
5. Did you plot univariate summaries of the data (histograms, density plots, boxplots)?

```
> p1 <- boston %>%
          ggplot(aes(crim, crim)) +
             geom_boxplot()
```

6. Did you consider correlations between variables (scatterplots)?

```
> boston %>% cor() %>% round(3)
# try the following
> round(cor(boston),2)
```

7. Did you check for outliers? What is an outlier? Single variable outlier or multivariate outlier? Why so important? What do we do? Throw away the samples? Man made sampling bias!



(a) Few noisy data. (b) Nonlinearly separable.

Figure 3.1: Data sets that are not linearly separable but are (a) linearly separable after discarding a few examples, or (b) separable by a more so phisticated curve.

Ref: Learning from data, A short course by Yaser Abu-Mostafa et al, p.79

8. Did you identify missing values?
9. Did you check the units of all data points to make sure they are in the right range?
10. Did you try to identify any errors or miscoding of variables?
11. Did you consider plotting on a log scale or any other transformation? Use log transforms for ratio measurements.
12. For large data sets, subsample before plotting
13. Use color and size to check for confounding

Workflow

- Do not let R to auto save your workspace, go to options in Rstudio, set "save workspace to .RData on exit" option to "never". This will enforce you to record every step you take while dealing with data and analysis.
- Create an RStudio project for each data analysis project.
- Keep data files there.
- Write your R scripts in an R markdown file. Keep scripts there.
- Save your outputs (plots and cleaned data) there.
- Only ever use relative paths, not absolute paths.
- DO NOT PRINT EVERYTHING!!!

Example Dataset - Ames

This problem presents a data set describing the sale of individual residential property in Ames, Iowa from 2006 to 2010. The data set contains 1460 observations and a large number of explanatory variables (nominal, ordinal, discrete, and continuous) involved in assessing home values.

With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this Kaggle competition challenges you to predict the final price of each home.

The full dataset as provided on the Kaggle website has been provided on Vula in the file ames.csv or in the following link: https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data