Basic EDA

Pramod Duvvuri

March 31, 2019

The below notes are written to accompany the book *Visualizing Data* by William Cleveland and the S670 class notes written by Prof. Dr. Brad Luen. Before learning Exploratory Data Analysis (EDA), one should be familiar with the basics of statistical concepts. One should also be familiar with Regression. The language of choice is **R** and the IDE is RStudio and we shall be using the *ggplot2* package of the *tidyverse* to plot, analyze and draw conclusions from the data we have. The references section will contain important resources that will aid you in understanding some tricky concepts that you shall encounter. Regarding the data, always pick datasets that have a lot of observations/rows, the minimum should be at least 100 observations.

1. Differences between CDA/EDA
2. What is EDA ?
    (a) Graphing
    (b) Fitting
3. The need for EDA
4. **Univariate Data** Single Measurement of a Quantitative Variable
    (a) Histogram
        i. Number of Bins
        ii. Binwidth
    (b) Density Plot (Frequency Polygon)
    (c) Boxplot
    (d) ECDF
    (e) Normal QQ Plot
    (f) Tukey Mean difference Plot
    (g) Additive Shift
    (h) Fitting a linear model
    (i) Residual Fitted Spread Plot
    (j) Skewness
    (k) Monotone Spread
    (l) Transformations
        i. Log Transform (log2/log10)
        ii. Power Transform
    (m) Spread Location Plot
5. **Bivariate Data** - Paired Measurements of Two Quantitative Variables
    (a) Correlation (Spearman/Kendall/Pearson)
    (b) Scatter Plot
    (c) loess curve (Parameters)
        i. α

ii. λ

(d) Curve with Confidence Intervals

(e) Skewed Data

i. Leptokurtic (Less Area in the Tails)

ii. Platykurtic (More Area in the Tails)

(f) Robust Fits

(g) Sliced Distribution Plots

References

1. http://docs.statwing.com/interpreting-residual-plots-to-improve-your-regression/

ggplot2 functions

```
ECDF - stat_ecdf()
Histogram - geom_histogram()
Density Plot - geom_density()
Boxplot - geom_boxplot()
Quantile Plot - stat_qq()
Facet Grid - facet_grid()
Facet Wrap - facet_wrap() \\ m x n display
Scatter Plot - geom_point()
Line - geom_abline()
QQ Plot - qqplot() \\ Base R function
Flip Axes - coord_flip()
Picking Colors - color()
Details in Histogram - scale_x_continuous()/scale_y_continuous()
To Set Axis Limits - coord_cartesian()
```