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.

1. Differences between CDA/EDA
2. What is EDA ?
   (a) Graphing
   (b) Fitting
3. The need for EDA
4. Univariate Data (singer data in lattice package)
   (a) Histogram
   (b) Density Plot
   (c) Boxplot
   (d) ECDF
   (e) Normal QQ Plot
   (f) Tukey Mean difference Plot
   (g) Additive Shift
   (h) Fitting a linear model
5. Bivariate Data
   (a) Scatter Plot

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()
```