LOAN DATA FROM PROSPER

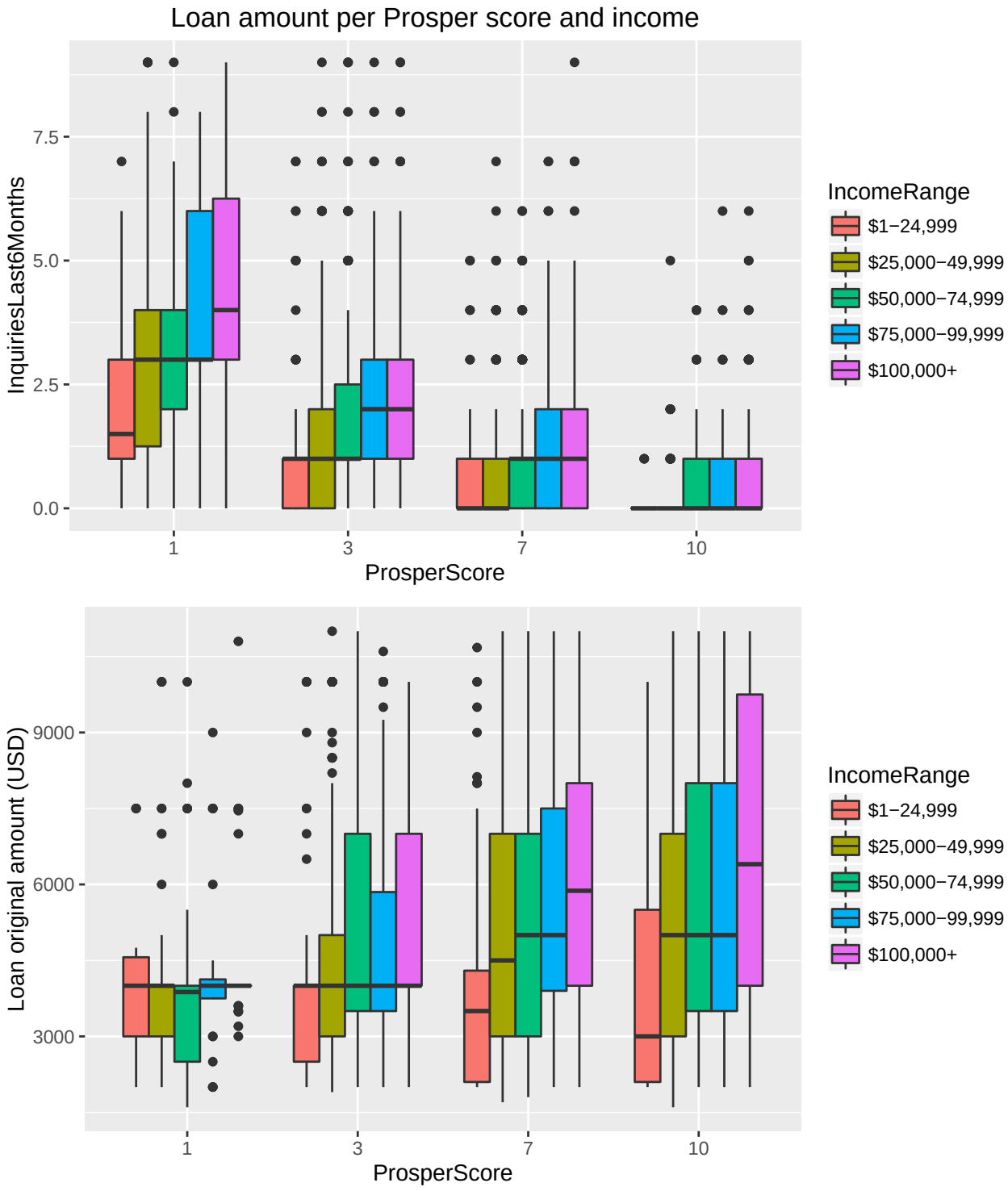
by *Paul F. Seke*

==========

Final Plots and Summary

After listing and exploring variables names, only the variables juged of interest for the study were extracted from the dataframe of 81 variables obtained from Proper CSV file (to improve memory use and analysis speed). Data were explored in terms of type (integer, boolean...) and quality (sufficient 'non missing' data). Where necessary, new variables obtained from transformation of pre-existing ones were created (using functions and iterations). Then data were proceeded for univariate, bivariate and multivariate plottings. Histograms, scatter plots, bar graphs, and line graphs were used, faceted and colored by categorical variables like home ownership, evidence of stated income, loan term, and Prosper score. Various methods were used to improve the quality of graphs and the detection of the information contained in the data, including: decreasing overoplotting (using alpha factor, jitters...); tranforming axes using the *sqrt* function in a loss less fashion (using *coord_cartesian* function); removing extreme values (e.g. by plotting data between quantiles 0.1-0.5 and 0.95-0.99); creating new categorical variables from continuous variables or categorical variables with either more general information (e.g. extraction of the delinquency status from the more general categorical variable "LoanStatus") or with numerous non-organized values (such as the categorical variable "ListingCategory" that had more than 20 possible values). As appropriate, proportions and means were determined for more accurate description of the data.

Plot One

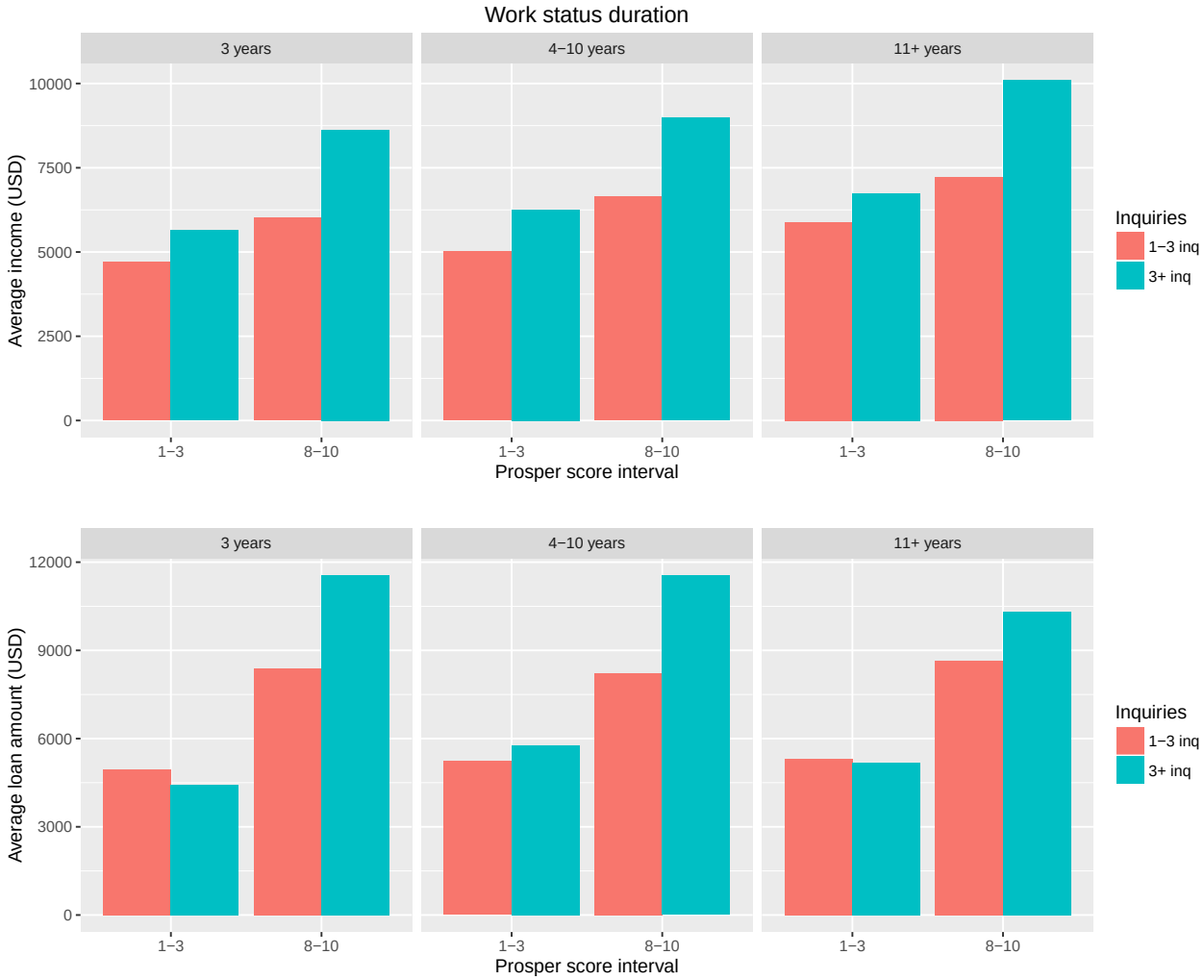


Description One

The boxplot of inquiries in the last 6 months in function of Prosper score and colored by income range (upper) suggested that people with high Prosper score (less risk) were more likely to be funded, including those with

very high incomes (> USD 100,000). The combination of that score and of a high monthly income appeared to be associated with funding at first attempt. The boxplot of loan amount in function of Prosper score and colored by income range (lower) suggested that people with both high Prosper score and high incomes were more likely to be funded for high amounts, while people with low Prosper score will not be funded for more than about USD 5,000. These observations summarize well the previous graphs and call for quantitative analysis of the impact of the income range, loan amount and Prosper score (that appeared as the strongest determinant for a loan in our study).

Plot Two



Description Two

The average monthly income in function of Prosper score, faceted by the work status duration and colored by the number of inquiries in the last 6 months further suggested that people with low Prosper score were not funded for more than USD 5,000. More specifically:
(*i*) for people with high Prosper score who had:
- 3 years or less work status duration: people with high Prosper score who were funded at first inquiry had an average salary of (± SEM) USD 5,901 ± 107 against USD 8,373 ± 1,045 for those who had to ask more times (t-test, p-value = 0.0247*);*

- *4 to 10 years work status duration: USD 6,614 ± 119 against USD 8,979 ± 895 (t-test, p-value = 0.0115)*;
- more than 11 years: USD 7,216 ± 136 against USD 10,091 ± 1,203 (t-test, p-value = 0.0231*);

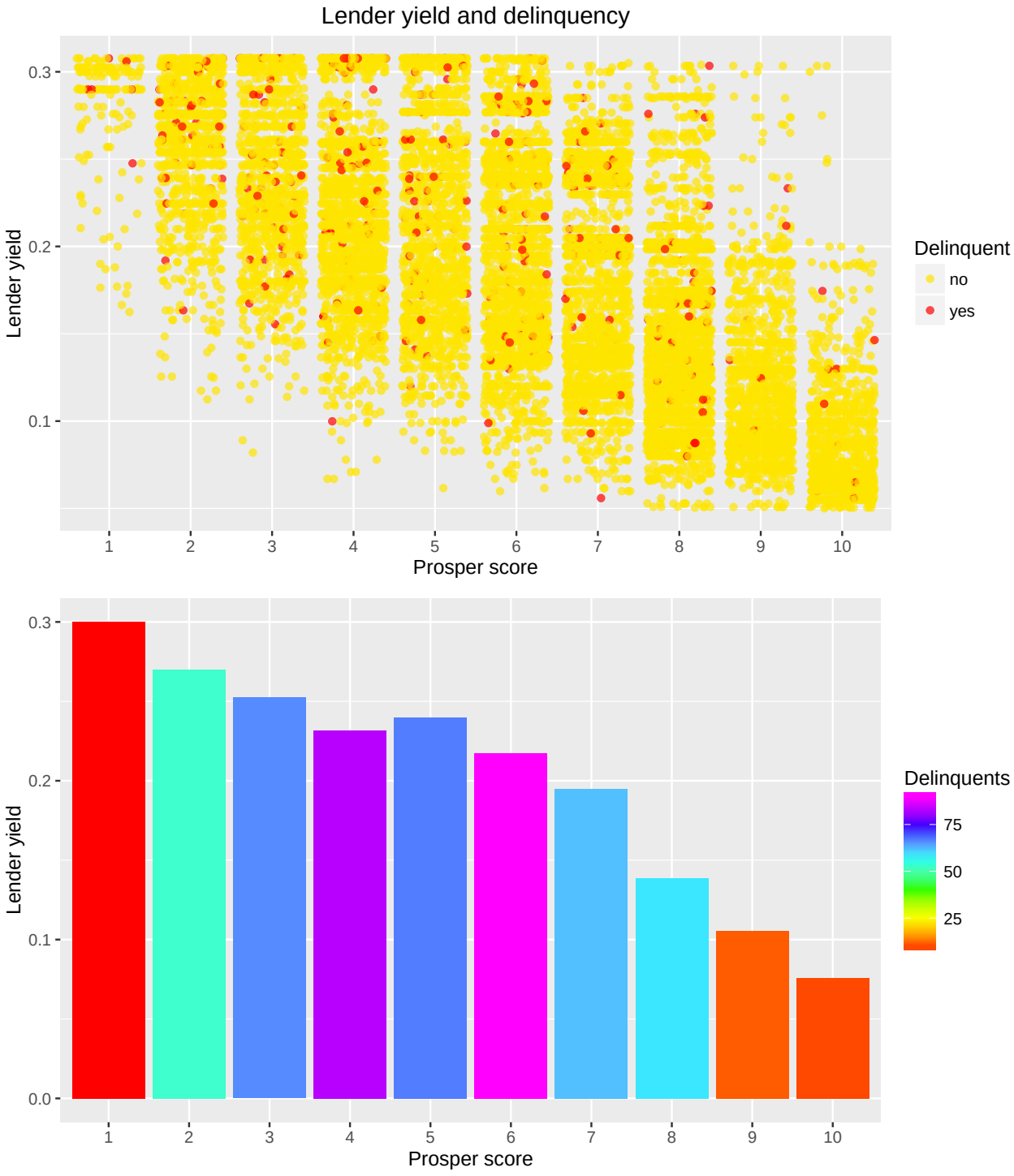
(*ii*) for people with low Prosper score who had:
- 3 years or less work status duration: USD 4,576 ± 125 against USD 5,507 ± 309 (t-test, p-value = 0.0059*);*
- *4 to 10 years work status duration: USD 4,951 ± 109 against USD 6,251 ± 326 (t-test, p-value = 0.0002)*;
- more than 11 years: USD 5,862 ± 158 against USD 6,745 ± 287 (t-test, p-value = 0.0076*).

Surprisingly, it appears that most people not funded at first attempt had high Prosper score and a good income. Analysing the average loan income provided a possible explanation, as (*iii*) people with high Prosper score who had:
- 3 years or less work status duration asked for USD 8,318 ± 171 against USD 11,564 ± 1,457 for those who had to ask more times (t-test, p-value = 0.0516);
- 4 to 10 years work status duration: USD 8,235 ± 145 against USD 11,544 ± 1,061 (t-test, p-value = 0.00308*);
- more than 11 years: USD 8,656 ± 170 against USD 10,302 ± 1,304 (t-test, p-value = 0.2178);

People with high Prosper score also had high income, but some were not funded at the first attempt as they were asking for high amounts.
Therefore, people with high Prosper score and high income were not funded at the first attempt probably because they were asking for high amounts.

(*iv*) No significant difference in loan amounts was observed between people with low scores inquiring once (USD 5,162 ± 71) or many times (USD 5,251 ± 165, p-value = 0.5982), confirming the graphical observation that people with low Prosper score were not funded for more than USD 5,000 despite their Prosper score, monthly income or work status duration.

Plot Three



Description Three

Quantitative analysis invalidated our assumption that delinquent were common in the high risk category (only 8 over 509 = 1.57%). They appeared instead to be common in the central interval of Prosper score normal distribution (368 over 509 = 72.3% between Prosper scores 3 and 7). On the other hand, quantitative

analysis confirmed our hypothesis from qualitative observation that Yielders gain increased with decreasing Prosper score (negative correlation). The average lender yield for Prosper score 1 was the highest (0.291) and Prosper score 10 yield was the lowest (0.198, p-value against score 1 < 2.2e-16).

Analysis

Total inquiries' and the last 6 months inquiries' statistical populations were significantly different in all filling categories. People applying for loans many times in the last 6 months also applied many time in the previous years, probably for business purpose, I hypothesized. The last 6 months' distribution had more new customers than returning customers, possibly indicating that at the time of data pulling Prosper was attracting new borrowers. On the other hand, the proportion of people filling for up to the 5th of the total amount (2 or 3 times in 6 months and 2 to 6 in total) was the highest in term of total inquiry, but only the second more common in the last 6 months' inquiry distribution, suggesting that the borrowers of Prosper tended to stay and keep on borrowing money. A limitation of the current dataset for our study is the non availability of the reasons why a client (on the basis of unique identifier) borrow money each time he/she comes (could be for different reasons each time. e.g. Personal, medical, business...). It could be interesting for future studies in this direction to retrieve data about the filling category of borrower all the previous times they applied for a loan an whether they got funded or not. Then, the real proportion of people funded in each category and after how money attempts could be accurately determined.
Nonetheless, in order to analyze loans for individual and familial motivations, we excluded people who applied for business at the time the credit profile was pulled. Although home ownership, monthly income, and the employment status duration correlated roughly with the loan funding, term and amount, Prosper risk score appeared to have a far stronger correlation. Quantitative analysis confirmed my qualitative observations. In addition, they also confirmed that delinquency was relatively scarce, but more common among people with Prosper score between 3 and 7, as the income per Prosper score followed a normal distribution. Yielders had a positive return in investment, despite delinquency, particularly for people investing on borrowers with higher risk (low Prosper score). From these results it appears to me that Prosper succeeded in its objective of providing a more humane approach to banking than classical financial institution as customer seem to remain loyal and new customers are attracted. It is also a successful business, as yielders have a return in investment between 19% to 30%, which is far better than the typical interests produced via banks. Future studies may address how the Prosper score is determined and how to improve it for borrowers providing better yield and less delinquency (currently mainly in the low score) to have better scores, and for the category with most delinquent to have a lower score (indicating more risk).

REFERENCES

docs.ggplot2.org - colour_fill_alpha
docs.ggplot2.org - current-theme docs.ggplot2.org - facet_grid
docs.ggplot2.org - geom_bar
docs.ggplot2.org - scale_brewer
docs.ggplot2.org- set-theme
docs.ggplot2.org - themes
en.wikipedia.org - peer-to-peer_lending
en.wikipedia.org - revolving credit
courses.statistics.com - R2prop
rmarkdown.rstudio.com - authoring basics
rmarkdown.rstudio.com - rcodechunks
stackoverflow.com - adding-x-and-y-axis-labels-in-ggplot2
stackoverflow.com - adjust color

stackoverflow.com - create a discrete color palette
stackoverflow.com - how-to-plot-one-variable-in-ggplot
stackoverflow.com - how-drop-data-frame-columns-by-name
stackoverflow.com - grouped-and-stacked-barplot
stackoverflow.com - creating-a-data-frame-from-two-vectors
stackoverflow.com - ggplot2-assigning-colours-to-a-factor
stackoverflow.com - ggplot-how-to-change-facet-labels
stackoverflow.com - how-do-i-manually-change-the-key-labels-in-a-legend-in-ggplot2
stackoverflow.com - how-to-plot-one-variable-in-ggplot
stackoverflow.com - how-to-assign-colors-to-categorical-variables-in-ggplot2
stackoverflow.com - how-to-sort-a-dataframe-by-columns
stackoverflow.com - plot-two-graphs-in-same-plot-in-r
stackoverflow.com - remove-all-of-x-axis-labels-in-ggplot
stackoverflow.com - remove-multiple-objects
stackoverflow.com - simplest-way-to-do-grouped-barplot
stackoverflow.com - turning-off-some-legends
stat.ethz.ch/R-manual
zevross.com - beautiful-plotting-in-r-a-ggplot2-cheatsheet
www.ats.ucla.edu - intro_function
www.cookbook-r.com - graphs colors
www.cookbook-r.com - Plotting_distributions
www.datacamp.com - tutorial-on-loops-in-r
www.fundingcircle
www.lendingmemo.com - lending-club-vs-prosper
www.myfico.com - crediteducation
www.programiz.com - r-if-else-statement
www.prosper.com - landing
www.r-bloggers.com - comparison-of-two-proportions
www.r-bloggers.com - one-way-analysis-of-variance
r-bloggers.com - from-continuous-to-categorical
www.r-tutor.com - two-population-proportions
www.statmethods.net - graphs/bar
www.statmethods.net - ttest
www.sthda.com - be-awesome-in-ggplot2
www.sthda.com - ggplot2-facet-split-a-plot-into-a-matrix-of-panels
www.sthda.com - ggplot2-colors
www.theanalysisfactor.com - r-tutorial-13
