Capstone Project: Data Wrangling Report

Sergey Mouzykin

November 2, 2017

Your raw data needs to be cleaned and wrangled into a form you can analyze. Here are a few suggested steps:

* Clean up your column names to be simple, short and descriptive
* For each column:
  + Check for missing values and decide what you want to do about them.
  + Make sure the values in each column make sense. If you find values that don't, decide what you want to do about those.
  + Look for outliers (values that are too small or too large). Do they make sense? Do you want to keep them in the data set?
  + Discuss with your mentor about other data wrangling steps you might need to perform for your specific problem and implement those.
  + Save your cleaned up and transformed data set.

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.3.3

library(readr)

## Warning: package 'readr' was built under R version 3.3.3

library(tidyr)

## Warning: package 'tidyr' was built under R version 3.3.3

library(dplyr)

## Warning: package 'dplyr' was built under R version 3.3.3

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

beer\_reviews <- read\_csv("beer\_reviews\_original.csv")

## Parsed with column specification:  
## cols(  
## brewery\_id = col\_integer(),  
## brewery\_name = col\_character(),  
## review\_time = col\_integer(),  
## review\_overall = col\_double(),  
## review\_aroma = col\_double(),  
## review\_appearance = col\_double(),  
## review\_profilename = col\_character(),  
## beer\_style = col\_character(),  
## review\_palate = col\_double(),  
## review\_taste = col\_double(),  
## beer\_name = col\_character(),  
## beer\_abv = col\_double(),  
## beer\_beerid = col\_integer()  
## )

beer\_reviews = as\_tibble(beer\_reviews)  
glimpse(beer\_reviews)

## Observations: 1,586,614  
## Variables: 13  
## $ brewery\_id <int> 10325, 10325, 10325, 10325, 1075, 1075, 107...  
## $ brewery\_name <chr> "Vecchio Birraio", "Vecchio Birraio", "Vecc...  
## $ review\_time <int> 1234817823, 1235915097, 1235916604, 1234725...  
## $ review\_overall <dbl> 1.5, 3.0, 3.0, 3.0, 4.0, 3.0, 3.5, 3.0, 4.0...  
## $ review\_aroma <dbl> 2.0, 2.5, 2.5, 3.0, 4.5, 3.5, 3.5, 2.5, 3.0...  
## $ review\_appearance <dbl> 2.5, 3.0, 3.0, 3.5, 4.0, 3.5, 3.5, 3.5, 3.5...  
## $ review\_profilename <chr> "stcules", "stcules", "stcules", "stcules",...  
## $ beer\_style <chr> "Hefeweizen", "English Strong Ale", "Foreig...  
## $ review\_palate <dbl> 1.5, 3.0, 3.0, 2.5, 4.0, 3.0, 4.0, 2.0, 3.5...  
## $ review\_taste <dbl> 1.5, 3.0, 3.0, 3.0, 4.5, 3.5, 4.0, 3.5, 4.0...  
## $ beer\_name <chr> "Sausa Weizen", "Red Moon", "Black Horse Bl...  
## $ beer\_abv <dbl> 5.0, 6.2, 6.5, 5.0, 7.7, 4.7, 4.7, 4.7, 4.7...  
## $ beer\_beerid <int> 47986, 48213, 48215, 47969, 64883, 52159, 5...

### LOWER CASE

Change all text to lower case Columns to change: brewery name, beer name, profile name, beer style

### RENAME COLUMNS

|  |  |
| --- | --- |
| Old Name | New Name |
| review\_overall | overall |
| review\_aroma | aroma |
| review\_appearance | appearance |
| review\_palate | palate |
| review\_taste | taste |
| review\_profilename | profile\_name |
| brewery\_name | *no change* |
| brewery\_id | *no change* |
| beer\_style | *no change* |
| beer\_name | *no change* |
| beer\_abv | *no change* |
| beer\_id | *no change* |
| review\_time | *no change* |

## Including Plots

You can also embed plots, for example:



Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code that generated the plot.