A3

James Tran

3/27/2018

# Assignment A3

The purpose of this assignment is to look at the categorical determinants of the value of a diamond based on the diamonds dataframe in the package ggplot2.

## Problem 1

First, load the libraries you need .Create a 10% random sample of the original dataframe, called lild. In this dataframe, create the variable price per carat (ppc). Keep only the variables, price, carat, ppc, cut color and clarity. Hopefully, by thinking of ppc as our target variable, the influence of carat has been removed.

Do a glimpse and a summary of lild to make sure everything has gone well.

library(tidyverse)

## -- Attaching packages --------------------------------------------------------------- tidyverse 1.2.1 --

## v ggplot2 2.2.1 v purrr 0.2.4  
## v tibble 1.4.2 v dplyr 0.7.4  
## v tidyr 0.7.2 v stringr 1.2.0  
## v readr 1.1.1 v forcats 0.2.0

## Warning: package 'ggplot2' was built under R version 3.4.4

## -- Conflicts ------------------------------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

diamonds %>%  
 sample\_frac(size=.1) %>%   
 mutate(ppc = price/carat) %>%   
 select(price, carat, ppc, cut,   
 color,clarity) ->  
 lild  
glimpse(lild)

## Observations: 5,394  
## Variables: 6  
## $ price <int> 4417, 2534, 8074, 1919, 3372, 14768, 2439, 2792, 1298,...  
## $ carat <dbl> 1.01, 0.81, 1.51, 0.52, 0.76, 2.01, 0.81, 0.71, 0.73, ...  
## $ ppc <dbl> 4373.267, 3128.395, 5347.020, 3690.385, 4436.842, 7347...  
## $ cut <ord> Fair, Good, Premium, Ideal, Ideal, Very Good, Ideal, P...  
## $ color <ord> G, H, H, D, G, H, I, E, G, D, E, D, D, G, F, H, I, E, ...  
## $ clarity <ord> SI1, SI1, SI2, VS2, VS1, SI2, SI1, SI1, I1, VS2, VVS2,...

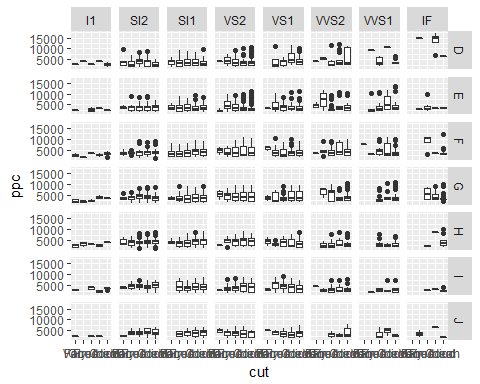
summary(lild)

## price carat ppc cut   
## Min. : 326 Min. :0.2200 Min. : 1152 Fair : 138   
## 1st Qu.: 936 1st Qu.:0.4000 1st Qu.: 2466 Good : 455   
## Median : 2386 Median :0.7000 Median : 3490 Very Good:1252   
## Mean : 3906 Mean :0.7958 Mean : 3996 Premium :1410   
## 3rd Qu.: 5415 3rd Qu.:1.0500 3rd Qu.: 4950 Ideal :2139   
## Max. :18797 Max. :4.0000 Max. :17083   
##   
## color clarity   
## D: 680 SI1 :1305   
## E: 976 VS2 :1243   
## F: 943 SI2 : 909   
## G:1113 VS1 : 809   
## H: 801 VVS2 : 535   
## I: 605 VVS1 : 377   
## J: 276 (Other): 216

## Problem 2

Now make an initial attempt at a visualization of how the three categorical variables influence ppc. Use facet\_grid to incorporate two of the three categorical variables. Map the third categorical variable to the horizontal axis of the cells in the grid. Choose the first of the two methods we demonstrated in class (geom\_boxplor() ) to show the relationship of the third categorical variable to the target ppc.

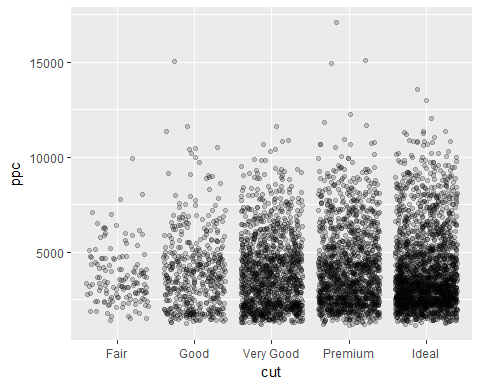
lild %>% ggplot(aes(x=cut,y=ppc)) + geom\_boxplot() +  
 facet\_grid(color~clarity)



## Problem 3

Basically repeat Problem 2, but change the type of grahic in each of the cells of the grid to use the second method we demonstrated in class (geom\_jitter() ). Choose a reasonable value of alpha.

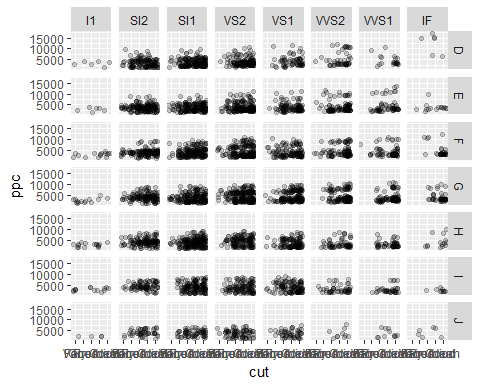
lild %>% ggplot(aes(x=cut,y=ppc)) + geom\_jitter(alpha=.2)



## Problem 4

Basically repeat Problem 3, but make a different choice on the roles of the three categorical variables.

lild %>% ggplot(aes(x=cut,y=ppc)) +   
 geom\_jitter(alpha=.2) +  
 facet\_grid(color~clarity)



## Problem 5

Compare what you see in these graphics. The main thing that I see within these graphics is that the the Ideal cut is one of the most expensive carats. Also There is more data related to the Very good, Premium, and Ideal cut’s over the Good and Fair.