Diamonds EDA Project 2 -Udacity Problem Set 4

Curtis ONeal

May 14, 2016

# In this problem set, you'll continue to explore the diamonds data set.

(<https://www.udacity.com/course/viewer#!/c-ud651/l-771308570/e-759808752/m-761758583>)

knitr::opts\_chunk$set(results =TRUE, fig.width=4, fig.height=4, echo=FALSE, warning=FALSE, message=FALSE)  
  
#load code library dependencies, and set up data set.  
library(ggplot2)  
library(gridExtra)  
library(magrittr)  
library(scales)  
library(plyr)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:plyr':  
##   
## arrange, count, desc, failwith, id, mutate, rename, summarise,  
## summarize

## The following object is masked from 'package:gridExtra':  
##   
## combine

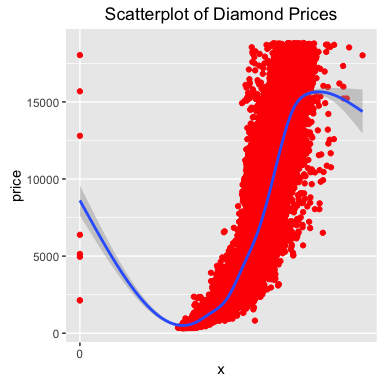
## The following objects are masked from 'package:stats':  
##   
## filter, lag

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

#Dataset diamonds is native to R, no need to load it.  
pf <- diamonds  
#Convert data set to pf shorter alias

Your first task is to create a scatterplot of price vs x using the ggplot syntax.

# Create a simple scatter plot of price vs x (a unit of diamond length).



## [1] "Price is positively correlated with x, a measure of a size dimension of the diamonds."

## [1] "Below are correlations related to excercise 3's questions in this section"

## [1] "Correlation of price to x:"

##   
## Pearson's product-moment correlation  
##   
## data: x and price  
## t = 440.16, df = 53938, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.8825835 0.8862594  
## sample estimates:  
## cor   
## 0.8844352

## [1] "Correlation of price to y:"

##   
## Pearson's product-moment correlation  
##   
## data: y and price  
## t = 401.14, df = 53938, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.8632867 0.8675241  
## sample estimates:  
## cor   
## 0.8654209

## [1] "Correlation of price to z:"

##   
## Pearson's product-moment correlation  
##   
## data: z and price  
## t = 393.6, df = 53938, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.8590541 0.8634131  
## sample estimates:  
## cor   
## 0.8612494

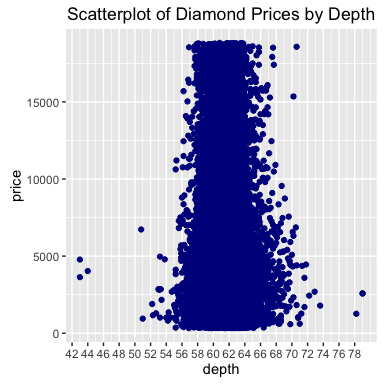
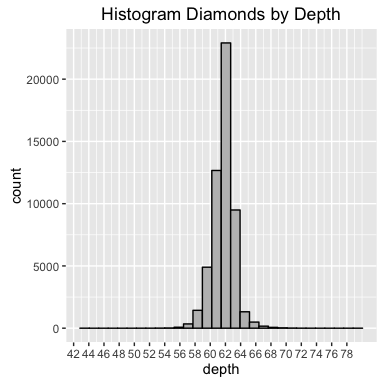
# A few other visualizations from excercises in this set.

Depth is a numerical unit of optical clarity for diamonds.

Create a simple scatter plot of price vs depth.

## [1] "Descriptive Stats of Depth"

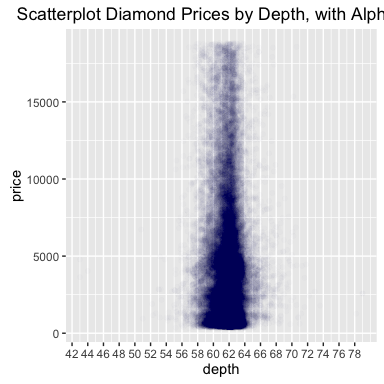
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 43.00 61.00 61.80 61.75 62.50 79.00



## [1] "Based on the Scatterplot of Prices vs Depth, nost diamonds are betwee 59 and 64 depth values."

Change the code to make the transparency of the points to be 1/100 of what they are now and mark the x-axis every 2 units.

# Scatterplot Diamond Prices by Depth with Alpha .01



## [1] "Correlation of price to depth: "

##   
## Pearson's product-moment correlation  
##   
## data: depth and price  
## t = -2.473, df = 53938, p-value = 0.0134  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.019084756 -0.002208537  
## sample estimates:  
## cor   
## -0.0106474

## [1] -0.01

## [1] "The correlation coefficient of price to depth is below .05, and seems too small as a predictor alone."

Create a scatterplot of price vs carat and omit the top 1% of price and carat values.

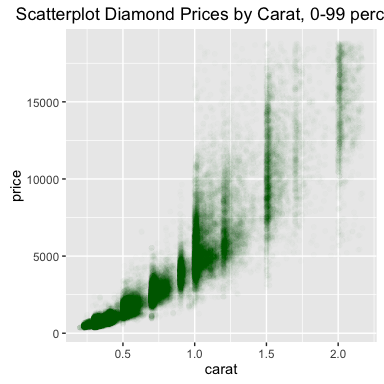
# Price Vs Carat, with top outliers removed

## [1] "Descriptive Statistics for Carat-"

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.2000 0.4000 0.7000 0.7979 1.0400 5.0100

## [1] "Define 99 Quantile for Carat-"

## 99%   
## 2.18

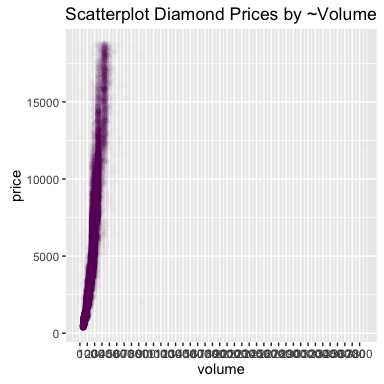


Create a scatterplot of price vs. volume (x \* y \* z). This is a very rough approximation for a diamond's volume.

# Define a rough volume estimage for diamond observations and plot it.

## [1] "Descriptive Statistics for Volume (defined as x\* y \* z)"

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 65.14 114.80 129.80 170.80 3841.00



## [1] "This should be zoomed, centered, and outliers removed."

## [1] "Descriptive Statistics of volume"

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 65.14 114.80 129.80 170.80 3841.00

## [1] "Defining bottom 1% of volume"

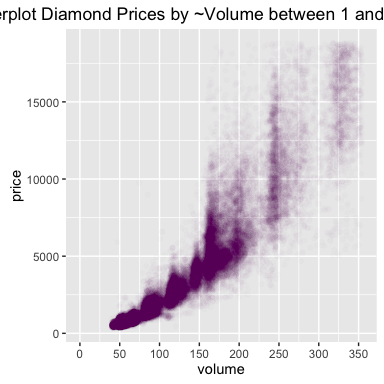
## 1%   
## 40.07633

## [1] "Defining 99 quantile of volume"

## 99%   
## 354.4266

## [1] "Defining Data betwee 1 and 99 quantile of volume"

## Source: local data frame [52,860 x 11]  
##   
## carat cut color clarity depth table price x y z  
## (dbl) (fctr) (fctr) (fctr) (dbl) (dbl) (int) (dbl) (dbl) (dbl)  
## 1 0.23 Very Good E VVS2 60.6 55 478 4.02 4.07 2.45  
## 2 0.23 Good G VVS1 56.2 60 395 4.14 4.19 2.34  
## 3 0.23 Very Good F VS2 59.4 57 402 4.06 4.09 2.42  
## 4 0.24 Premium H VVS2 60.7 58 554 4.07 4.04 2.46  
## 5 0.24 Ideal G VVS1 62.2 56 559 4.00 4.04 2.50  
## 6 0.24 Ideal G VVS1 62.0 55 559 4.01 4.03 2.49  
## 7 0.24 Very Good E SI1 59.9 58 571 4.06 4.10 2.44  
## 8 0.24 Ideal F VS1 62.7 54 572 3.99 4.02 2.51  
## 9 0.24 Premium E VS2 62.6 58 417 3.99 4.03 2.51  
## 10 0.24 Very Good E VS2 61.5 58 419 4.02 4.05 2.48  
## .. ... ... ... ... ... ... ... ... ... ...  
## Variables not shown: volume (dbl)



Did you notice some outliers? Some volumes are 0! There's an expensive diamond with a volume near 4000, and a cheaper diamond with a volume near 900. You can find out how many diamonds have 0 volume by using count(diamonds$volume == 0). The count() function comes with the plyr package.

# Volumes that are 0!

## [1] "The number of diamonds with volume = 0:"

## x freq  
## 1 FALSE 53920  
## 2 TRUE 20

What's the correlations of price and volume excluding diamonds with 0 volume, and over 800?

# Correlations as above

## [1] "The prices and volume excluding diamonds with 0 volume, and over 800 are a full vector of 53,917 numbers. Below is the desciptive statistical summary of these values:"

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 31.71 65.19 114.80 129.80 170.80 790.10

## [1] "The correlation of prices and volume, excluding diamonds with 0 volume, and over 800 is:"

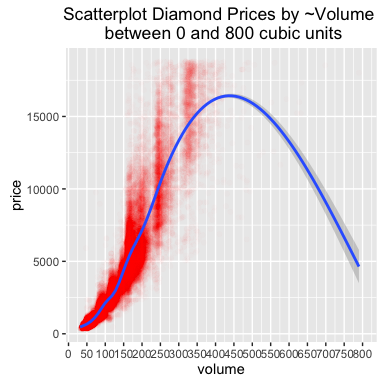
## [1] 0.92

INLINE VARIBLE TEST: 11

Subset the data to exclude diamonds with a volume greater than or equal to 800. Also, exclude diamonds with a volume of 0. Adjust the transparency of the points and add a linear model to the plot. Use geom\_smooth() for this.

We encourage you to think about this next question and to post your thoughts in the discussion section.

Do you think this would be a useful model to estimate the price of diamonds? Why or why not?



Use the function dplyr package to create a new data frame containing info on diamonds by clarity. Name the data frame diamondsByClarity

The data frame should contain the following variables in this order.

(1) mean\_price  
 (2) median\_price  
 (3) min\_price  
 (4) max\_price  
 (5) n

where n is the number of diamonds in each level of clarity.

# Create the DaimondsByClarity dataframe

## [1] "First few Rows of Diamonds"

## Source: local data frame [6 x 10]  
##   
## carat cut color clarity depth table price x y z  
## (dbl) (fctr) (fctr) (fctr) (dbl) (dbl) (int) (dbl) (dbl) (dbl)  
## 1 0.23 Ideal E SI2 61.5 55 326 3.95 3.98 2.43  
## 2 0.21 Premium E SI1 59.8 61 326 3.89 3.84 2.31  
## 3 0.23 Good E VS1 56.9 65 327 4.05 4.07 2.31  
## 4 0.29 Premium I VS2 62.4 58 334 4.20 4.23 2.63  
## 5 0.31 Good J SI2 63.3 58 335 4.34 4.35 2.75  
## 6 0.24 Very Good J VVS2 62.8 57 336 3.94 3.96 2.48

## [1] "Define Diamonds by Clarity"

## Source: local data frame [6 x 6]  
##   
## clarity mean\_price median\_price min\_price max\_price n  
## (fctr) (dbl) (dbl) (int) (int) (int)  
## 1 I1 3924.169 3344 345 18531 741  
## 2 SI2 5063.029 4072 326 18804 9194  
## 3 SI1 3996.001 2822 326 18818 13065  
## 4 VS2 3924.989 2054 334 18823 12258  
## 5 VS1 3839.455 2005 327 18795 8171  
## 6 VVS2 3283.737 1311 336 18768 5066

## [1] "Not sure about the Error -Error in n() : This function should not be called directly"

Note:Another warning: Version 0.4.0 of dplyr has a bug when using the median function on the summarize layer, depending on the nature of the data being summarized. You may need to cast the data as a numeric (float) type to get the expected results, e.g. median(as.numeric(var)).

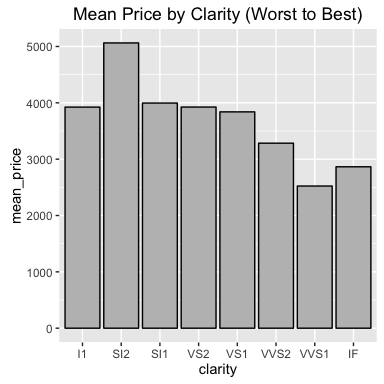
We’ve created summary data frames with the mean price by clarity and color. You can run the code in R to verify what data is in the variables diamonds\_mp\_by\_clarity and diamonds\_mp\_by\_color.

Your task is to write additional code to create two bar plots on one output image using the grid.arrange() function from the package gridExtra.

# Our diamonds\_mp\_by\_clarity and diamonds\_mp\_by\_color

## Source: local data frame [6 x 10]  
## Groups: clarity [5]  
##   
## carat cut color clarity depth table price x y z  
## (dbl) (fctr) (fctr) (fctr) (dbl) (dbl) (int) (dbl) (dbl) (dbl)  
## 1 0.23 Ideal E SI2 61.5 55 326 3.95 3.98 2.43  
## 2 0.21 Premium E SI1 59.8 61 326 3.89 3.84 2.31  
## 3 0.23 Good E VS1 56.9 65 327 4.05 4.07 2.31  
## 4 0.29 Premium I VS2 62.4 58 334 4.20 4.23 2.63  
## 5 0.31 Good J SI2 63.3 58 335 4.34 4.35 2.75  
## 6 0.24 Very Good J VVS2 62.8 57 336 3.94 3.96 2.48

## Source: local data frame [6 x 3]  
##   
## color mean\_price mean\_carat  
## (fctr) (dbl) (dbl)  
## 1 D 3169.954 0.6577948  
## 2 E 3076.752 0.6578667  
## 3 F 3724.886 0.7365385  
## 4 G 3999.136 0.7711902  
## 5 H 4486.669 0.9117991  
## 6 I 5091.875 1.0269273



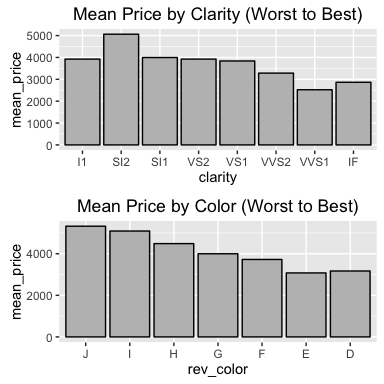
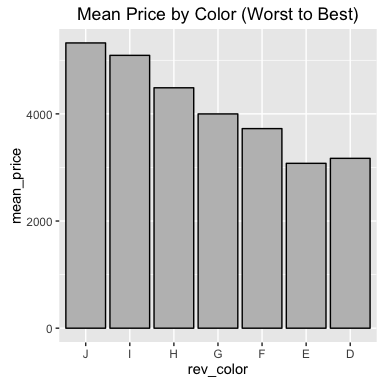
## [1] "For chart display purposes reversing Color factor order so Best is in the Right."

## [1] "original"

## [1] D E F G H I J  
## Levels: D < E < F < G < H < I < J

## [1] "reversed"

## [1] D E F G H I J  
## Levels: J < I < H < G < F < E < D



From help Diamonds:

color: diamond colour, from J (worst) to D (best)

clarity: a measurement of how clear the diamond is (I1 (worst), SI1, SI2, VS1, VS2, VVS1, VVS2, IF (best)

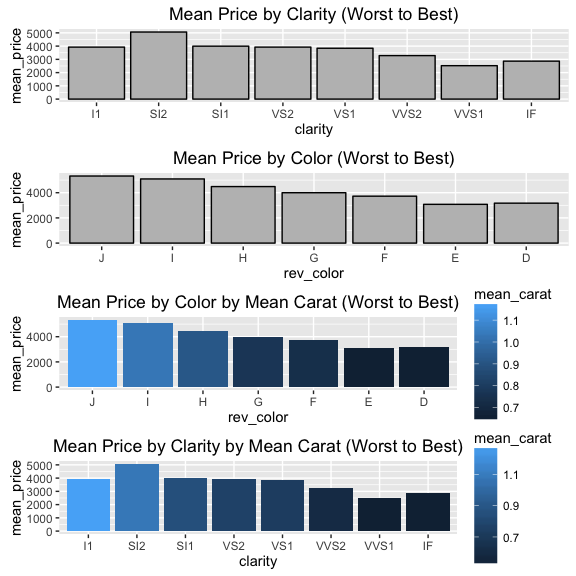
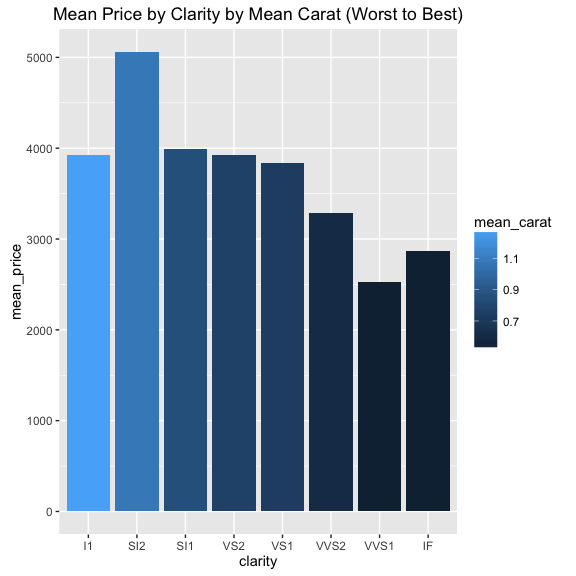
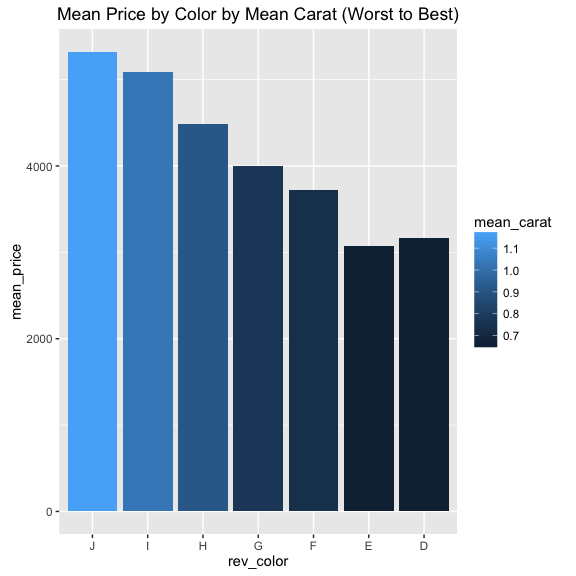
Reflecting on this, the plots are counter intuitive.

The dataset descriptions shows the best categroy of clarity is IF and the best category of color is D, but these values are the 2nd worst for average prices for both values by category. One would expect that the best cateogory would have the highest prices.

# Investiage pattern with price

## [1] "Table of Mean prices and Mean Carats by Color"

## Source: local data frame [7 x 4]  
##   
## color mean\_price mean\_carat rev\_color  
## (fctr) (dbl) (dbl) (fctr)  
## 1 D 3169.954 0.6577948 D  
## 2 E 3076.752 0.6578667 E  
## 3 F 3724.886 0.7365385 F  
## 4 G 3999.136 0.7711902 G  
## 5 H 4486.669 0.9117991 H  
## 6 I 5091.875 1.0269273 I  
## 7 J 5323.818 1.1621368 J



By adding average Carat size by clairty and color we see that the best diamonds of this class on average are smaller than the other grades of diamonds for clarity and color.

Since size is so strongly correlated with price, it would seem to be the dominant factor on average affecting the prices.

A separate examination of a subset of only the highest price diamonds would be needed to determine if these have the highest factos of color and clarity.