Case Study: Diamonds

Dr. Aijun Zhang
STAT3622 Data Visualization

26 September 2016

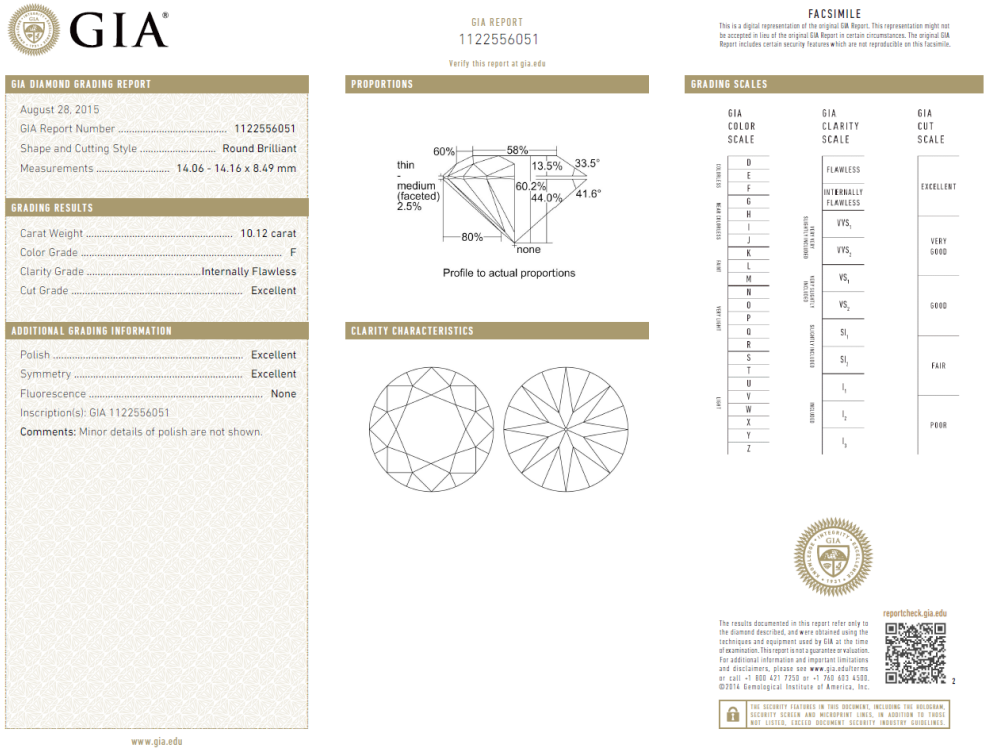
Department of 統計及精算學系
Statistics & Actuarial Science

Outline

1. A Brief Diamond Education
2. Data Manipulation with **dplyr**
3. Data Visualization with **ggplot2**
4. Not An End

GIA Diamond Grading Certificate

- When you may buy a diamond ring, you will get a report like



GIA

GIA DIAMOND GRADING REPORT

August 28, 2015
GIA Report Number 1122556051
Shape and Cutting Style ..... Round Brilliant
Measurements ..... 14.06 - 14.16 x 8.49 mm

GRADING RESULTS

Carat Weight ..... 10.12 carat
Color Grade ..... F
Clarity Grade ..... Internally Flawless
Cut Grade ..... Excellent

ADDITIONAL GRADING INFORMATION

Polish ..... Excellent
Symmetry ..... Excellent
Fluorescence ..... None
Inscription(s): GIA 1122556051
Comments: Minor details of polish are not shown.

www.gia.edu

GIA REPORT
1122556051
Verify this report at gia.edu

PROPORTIONS

Profile to actual proportions

CLARITY CHARACTERISTICS

FACSIMILE

This is a digital representation of the original GIA Report. This representation might not be accepted in lieu of the original GIA Report in certain circumstances. The original GIA Report includes certain security features which are not reproducible on this facsimile.

GRADING SCALES

reportcheck.gia.edu

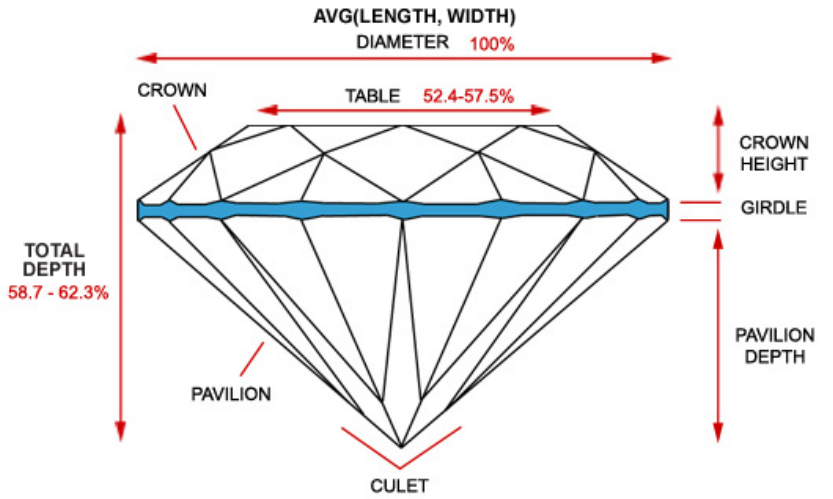
The results documented in this report refer only to the diamond described, and were obtained using the techniques and equipment used by GIA at the time of examination. This report is not a guarantee or valuation. For additional information and important limitations and disclaimers, please see www.gia.edu/terms or call +1 800 421 7250 or +1 760 603 4500. ©2014 Gemological Institute of America, Inc.

THE SECURITY FEATURES IN THIS DOCUMENT, INCLUDING THE HOLOGRAM, SECURITY SCREEN AND MICROPRINT LINES, IN ADDITION TO THOSE NOT LISTED, EXCEED DOCUMENT SECURITY INDUSTRY GUIDELINES.

- Read about the 4 C's (Carat, Cut, Color, Clarity)
- This 10.12 carat diamond costs ∼1MM USD at this link

Diamond Measurements

- Th cut grade is determined by {length (mm), width (mm), total depth(%) and table width(%), ... }.
- For example of ideal round cut diamond:



- A carat-size chart for standard 57 facets round brilliant cut:

Carat Weight	0.25	0.50	0.75	1.00	1.50	2.00	5.00
Round	4.1mm	5.2mm	5.8mm	6.5mm	7.4mm	8.2mm	11.1mm

This Case Study

- Consider real dataset of price and quality information of a large amount of diamonds
- Learn some basics of data manipulation (or data wrangling)
- Conduct exploratory data analysis with graphical tools we have discussed so far

Outline

1 A Brief Diamond Education

2 Data Manipulation with **dplyr**

3 Data Visualization with **ggplot2**

4 Not An End

The ‘diamonds‘ Data collected in 2008

```
## # A tibble: 53,940 × 10
##    carat       cut color clarity depth table price     x     y     z
##    <dbl>     <ord> <ord>   <ord> <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
## 7   0.24 Very Good     I    VVS1  62.3    57   336  3.95  3.98  2.47
## 8   0.26 Very Good     H     SI1  61.9    55   337  4.07  4.11  2.53
## 9   0.22      Fair     E     VS2  65.1    61   337  3.87  3.78  2.49
## 10  0.23 Very Good     H     VS1  59.4    61   338  4.00  4.05  2.39
## # ... with 53,930 more rows
```

Built in with ggplot2 package. Use `?diamonds` to check data details. But what is a ‘tibble‘?

R Packages for Data Wraggling

Besides **ggplot2()** for data visualization, Hadley Wickham has created a series of R packages for data wraggling, including

- **tidyr** for tidy data: observations in rows, variables in columns
- **tibble** for better ways to create, print and subset data frames
- **dplyr** for data manipulation $\Rightarrow$
- etc ...

Refer to the ggplot2 book Chapter 9 about Tidy Data; and Chapter 10 about **dplyr**.

Data Manipulation with **dplyr**

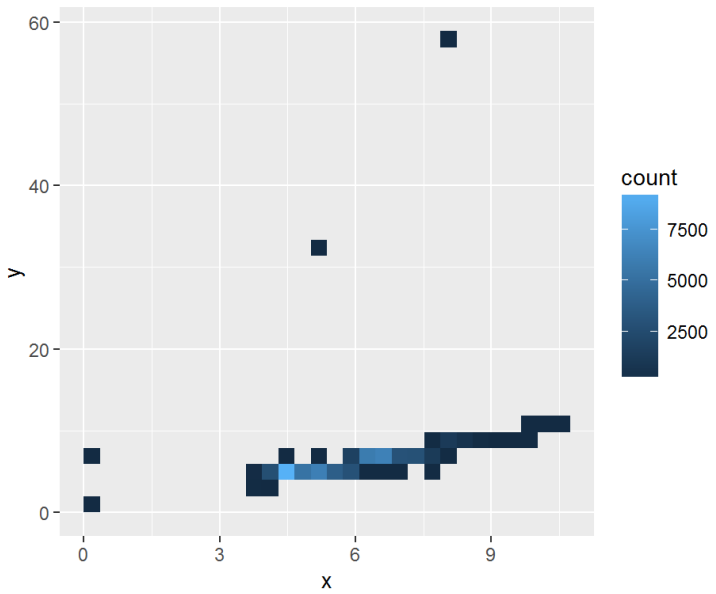
We will manipulate the ‘diamonds‘ data with the **dplyr** verbs:

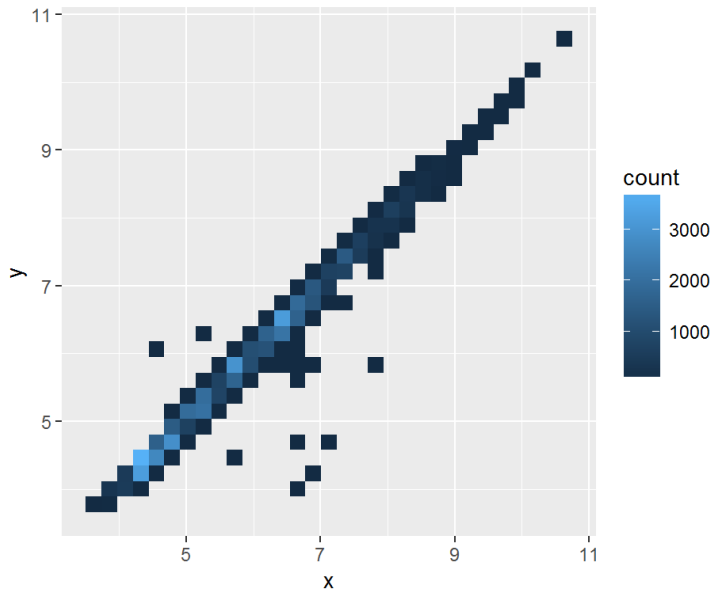
- **filter** to select the observations
- **mutate** to create new variables
- **group_by** to group variables
- **%>%** to code sequential operations

Other verbs: select, arrange, rename, _join, bind_, summarize, ...

dplyr::filter

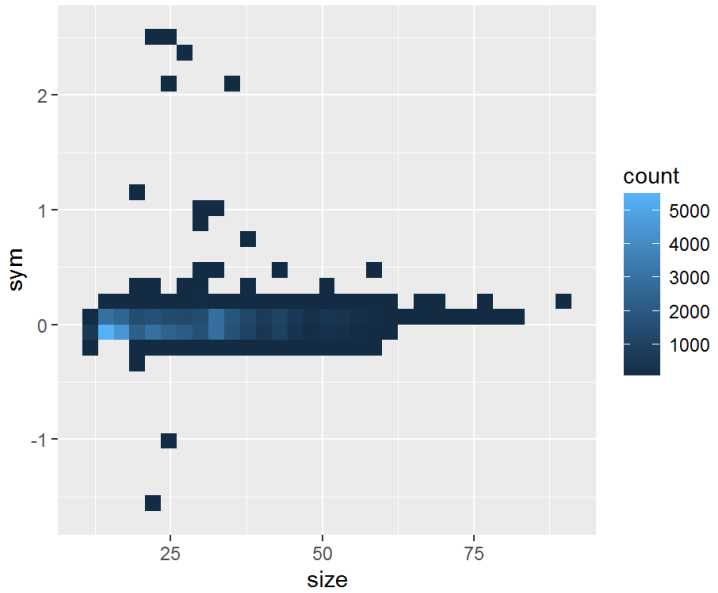
```
ggplot(diamonds, aes(x,y)) + geom_bin2d()
diamonds1 = filter(diamonds, x>0, y>0, y<20)
ggplot(diamonds1, aes(x,y)) + geom_bin2d()
```

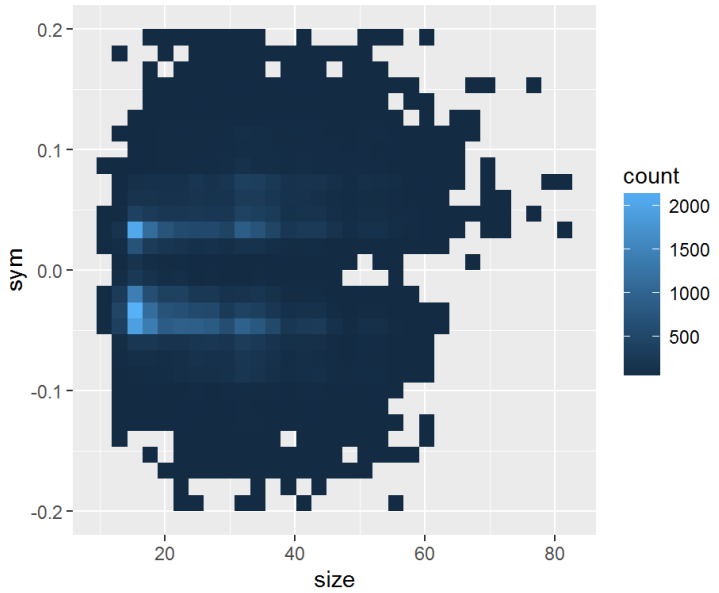




dplyr::mutate

```
diamonds2 = mutate(diamonds1, sym=x-y,
                   size=pi*((x+y)/2/2)^2)
ggplot(diamonds2, aes(size,sym)) + geom_bin2d()
diamonds3 = filter(diamonds2, abs(sym)<0.2)
ggplot(diamonds3, aes(size,sym)) + geom_bin2d()
```



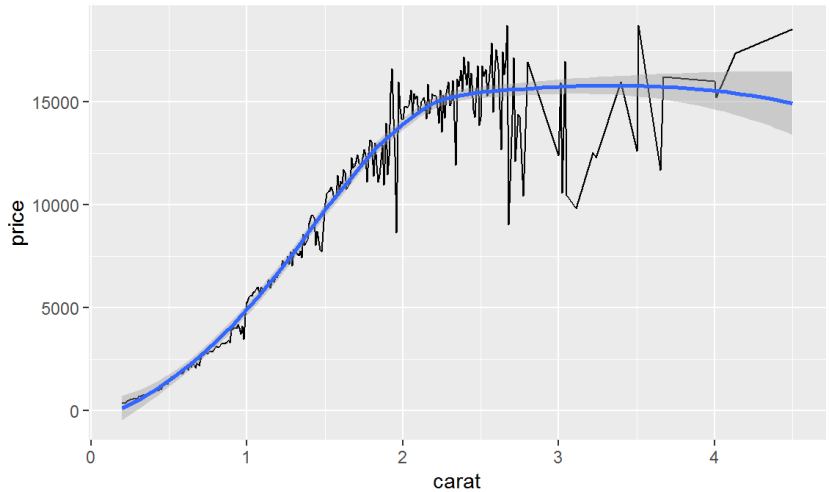


dplyr::group_by

```
tmp1 = group_by(diamonds3, clarity)
tmp2 = summarize(tmp1, price=mean(price))
> tmp2
# A tibble: 8 x 2
  clarity    price
    <ord>    <dbl>
1      I1 3943.831
2     SI2 5058.211
3     SI1 3995.493
4     VS2 3922.378
5     VS1 3835.650
6    VVS2 3283.985
7    VVS1 2515.867
8      IF 2857.964
```

dplyr::%>%

```
diamonds %>%
  filter(x>0, y>0, y<20) %>%
  mutate(sym=x-y, size=pi*((x+y)/2/2)^2) %>%
  filter(abs(sym)<0.2) %>%
  group_by(carat) %>%
  summarize(price=mean(price)) %>%
  ggplot(aes(carat, price)) +
  geom_line() +
  geom_smooth()
```



Outline

1. A Brief Diamond Education
2. Data Manipulation with **dplyr**
3. Data Visualization with **ggplot2**
4. Not An End

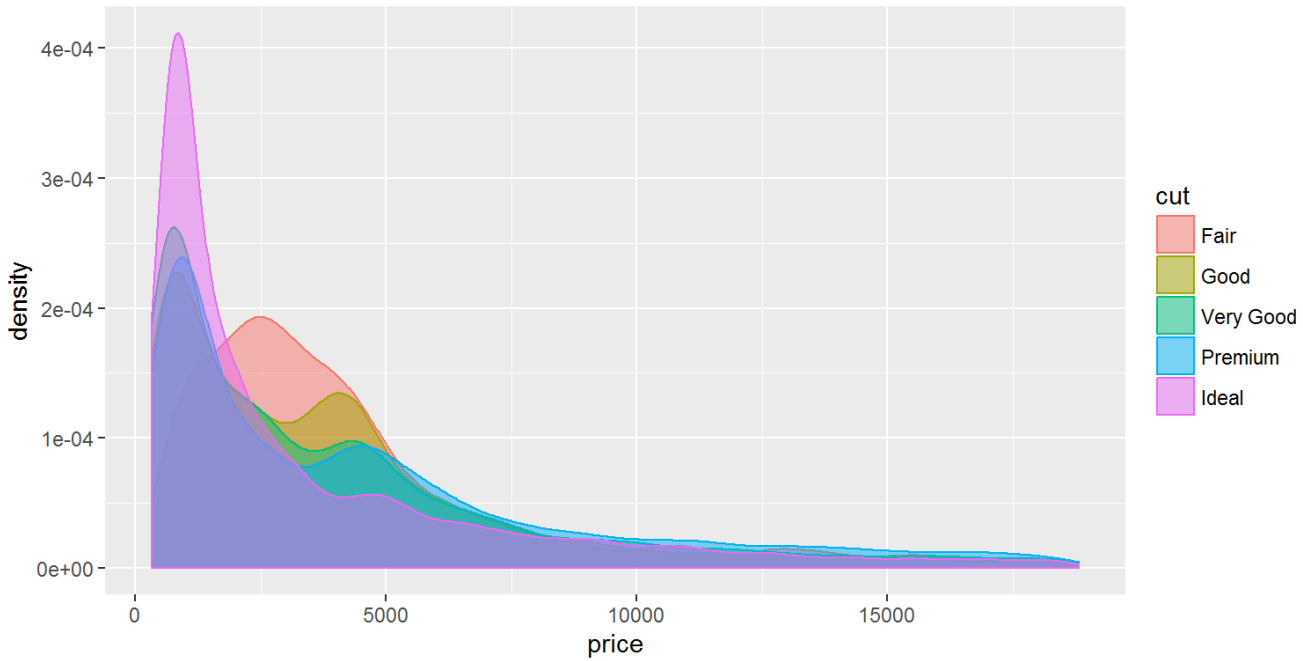
Data Visualization with **ggplot2**

Based on the diamonds dataset after removing certain outliers, we use **ggplot2** to demonstrate:

- Displaying distributions
- Dealing with overplotting
- Transforming variables

among others.

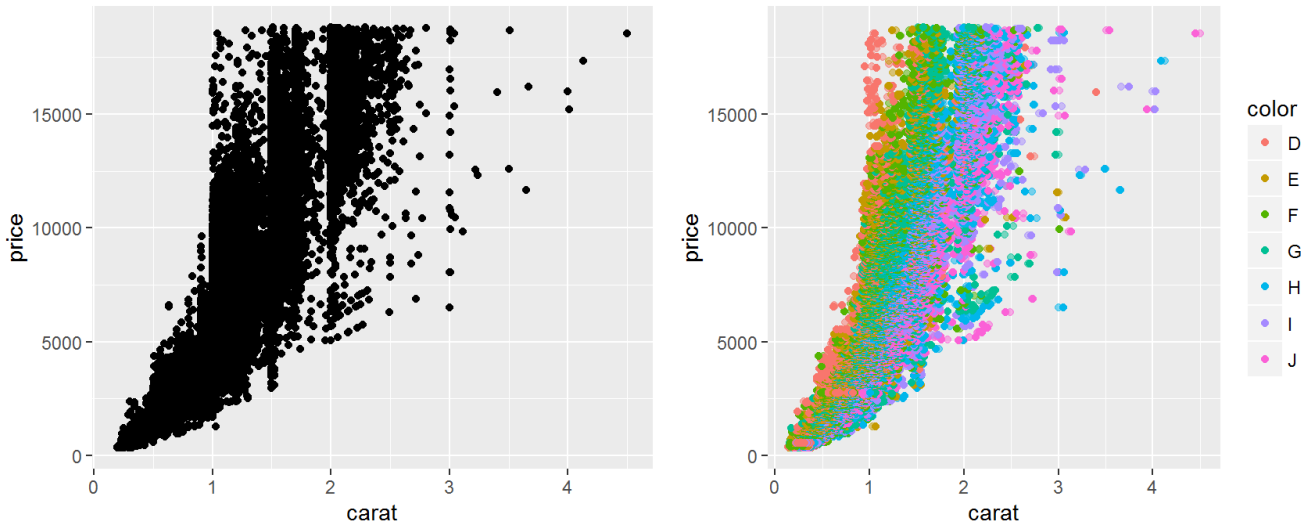
Displaying distributions



Do you find anything counter-intuitive?

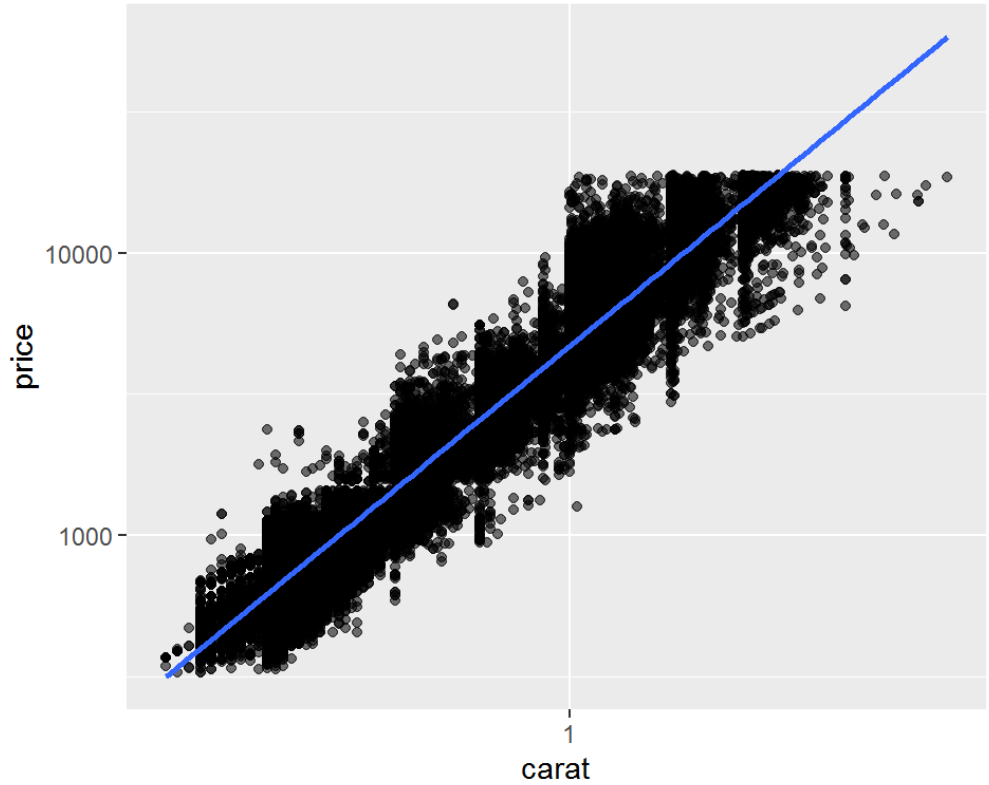
Dealing with overplotting

Use transparency (i.e. seting alpha) and geom_jitter():



Transforming variables

Take log-transform for both x and y results in linear relationship:



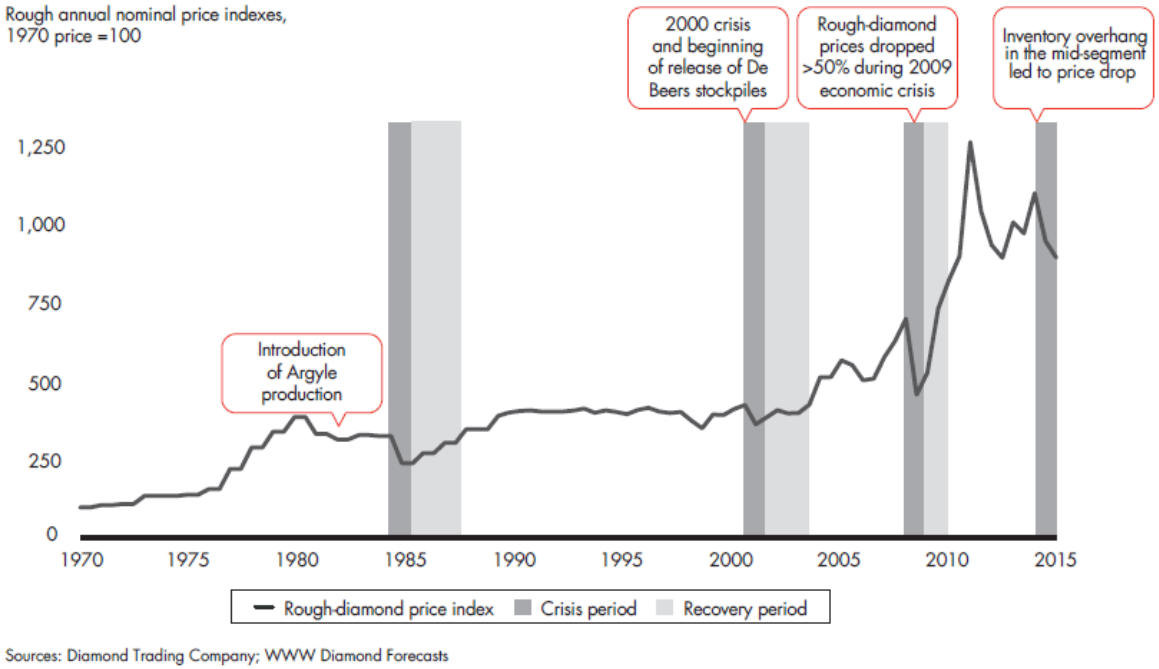
Can you explain why the top right points are cut off?

Outline

1. A Brief Diamond Education
2. Data Manipulation with **dplyr**
3. Data Visualization with **ggplot2**
4. Not An End

Further Understanding of Diamond Market

- Read an interesting blog post by Solomon Messing for an extensive analysis of diamonds dataset conducted in 2014.
- A historical view of global diamond index:



- Can you conduct an updated diamonds analytics with more complete dataset?