Introduction to Statistics and Statistical Measures

Adejumo Ridwan Suleiman

June 25 2022

Section 1

Introduction to Statistics and Statistical Measures

What is Statistics?

- Statistics is the collection, organizing and analysing of data.

Is Data Science Statistics in Disguise?

- Unlike Statistics, Data Science is an interdisciplinary field consisting of Mathematics, Statistics, Computer Science and Domain Knowledge.

Types of Data

- Data can be classified into two types
  - Based on Measurement scale
  - Based on Time Period

Based on Measurement Scale

- Qualitative Data
  - Nominal Data e.g sex
  - Ordinal Data e.g temperature level; High, Medium and Low
- Quantitative Data
  - Ratio e.g weight
  - Interval e.g temperature in degreee celsius

Based on Time Period

- Cross-Sectional Data e.g number of viewers for different youtube genres in the year 2021
- Time Series Data e.g number of viewers for Sport channels on youtube from the year 2014-Date.

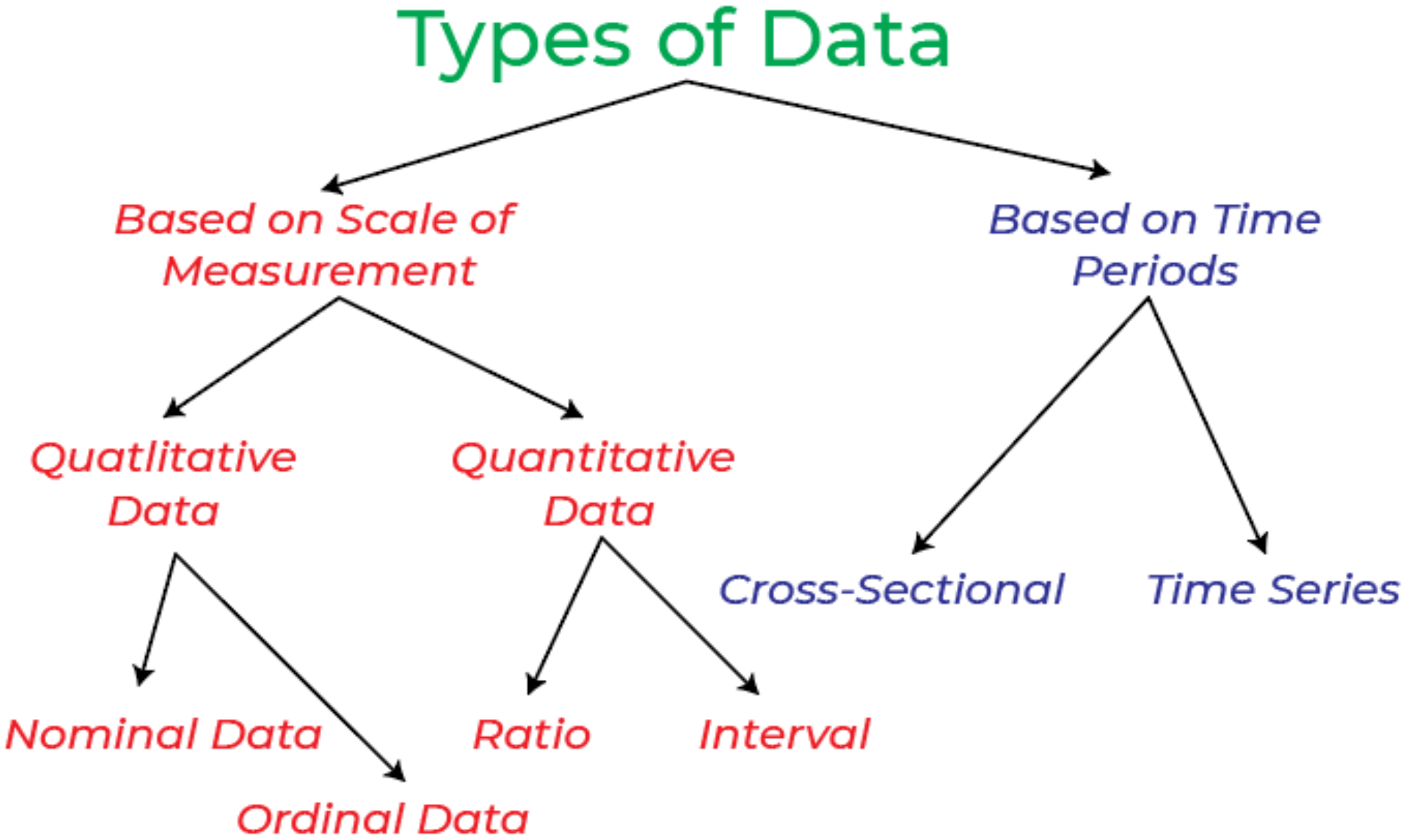


Figure 1: Types of Data

Rectangular or Structured Data

	carat	cut	color	clarity	depth	table	price	x	y	z
1	0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
2	0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
3	0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
4	0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
5	0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75
6	0.24	Very Good	J	VVS2	62.8	57.0	336	3.94	3.96	2.48
7	0.24	Very Good	I	VVS1	62.3	57.0	336	3.95	3.98	2.47
8	0.26	Very Good	H	SI1	61.9	55.0	337	4.07	4.11	2.53
9	0.22	Fair	E	VS2	65.1	61.0	337	3.87	3.78	2.49
10	0.23	Very Good	H	VS1	59.4	61.0	338	4.00	4.05	2.39
11	0.30	Good	J	SI1	64.0	55.0	339	4.25	4.28	2.73
12	0.23	Ideal	J	VS1	62.8	56.0	340	3.93	3.90	2.46
13	0.22	Premium	F	SI1	60.4	61.0	342	3.88	3.84	2.33
14	0.31	Ideal	J	SI2	62.2	54.0	344	4.35	4.37	2.71
15	0.20	Premium	E	SI2	60.2	62.0	345	3.79	3.75	2.27
16	0.32	Premium	E	I1	60.9	58.0	345	4.38	4.42	2.68
17	0.30	Ideal	I	SI2	62.0	54.0	348	4.31	4.34	2.68
18	0.30	Good	J	SI1	63.4	54.0	351	4.23	4.29	2.70

Measures of Central Tendency

Mean

- Sum of all values of observations divided by the number of observations
- Mathematically denoted as:
  $\bar{a} = \frac{\sum_i^n x_i}{n}$
- Sensitive to extreme or high values

Median

- Center of an ordered observations
- Also known as the middle of the observations.
- Not sensitive to extreme values

Mode

- Observation with the highest number of occurence.

Measures of Variation

Standard Deviation and Variance

- Measures how far an observation is from the mean
- Mathematically defined as:

$$s = \sqrt{\frac{\sum_i^n (x_i - \bar{x})}{n}}$$

- Variance is defined as the square of the standard deviation:
  $Variance = s^2$

Range

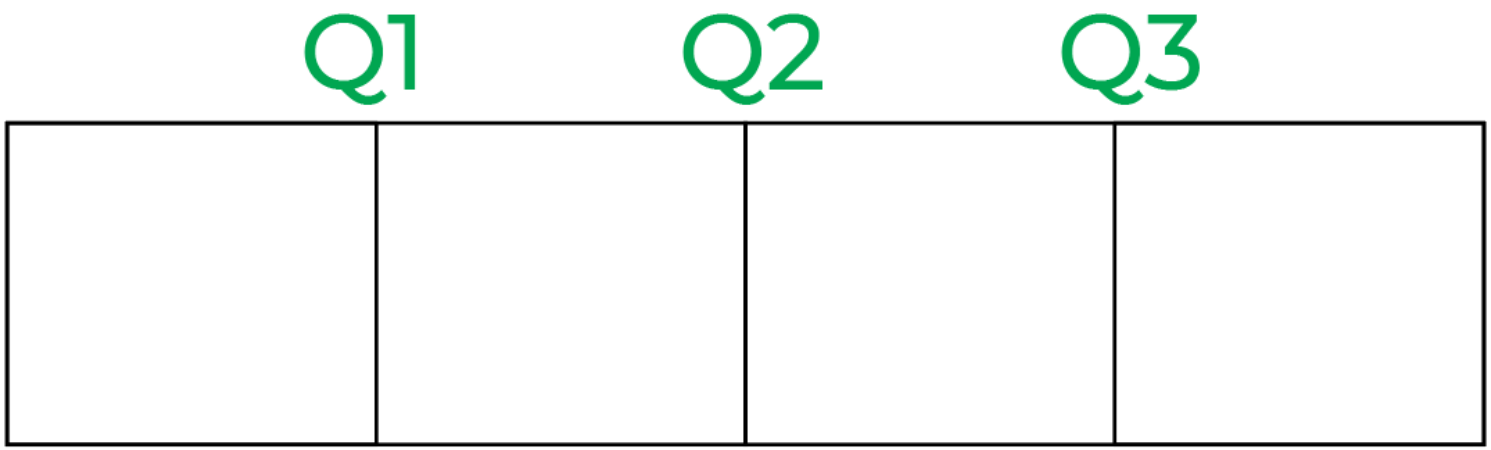
- Difference between the largest and smallest observations.
- Sensitive to extreme values

Percentiles

- Expressing the sorted observations in percentage
- Not sensitive to extreme values

Interquartile Range

- The interquartile range divides the observations into 4 equal part:
  - First Quartile: Q1
  - Second Quratile: Q2 (median)
  - Third Quratile: Q3



Summary Statistics of the Diamond Data Set

```
     carat                cut        color        clarity          depth      
 Min.   :0.2000   Fair      : 1610   D: 6775   SI1     :13065   Min.   :43.00  
 1st Qu.:0.4000   Good      : 4906   E: 9797   VS2     :12258   1st Qu.:61.00  
 Median :0.7000   Very Good:12082   F: 9542   SI2     : 9194   Median :61.80  
 Mean   :0.7979   Premium   :13791   G:11292   VS1     : 8171   Mean   :61.75  
 3rd Qu.:1.0400   Ideal     :21551   H: 8304   VVS2    : 5066   3rd Qu.:62.50  
 Max.   :5.0100                     I: 5422   VVS1    : 3655   Max.   :79.00  
                                    J: 2808   (Other): 2531                  
     table           price                x                y         
 Min.   :43.00   Min.   :  326   Min.   : 0.000   Min.   : 0.000  
 1st Qu.:56.00   1st Qu.:  950   1st Qu.: 4.710   1st Qu.: 4.720  
 Median :57.00   Median : 2401   Median : 5.700   Median : 5.710  
 Mean   :57.46   Mean   : 3933   Mean   : 5.731   Mean   : 5.735  
 3rd Qu.:59.00   3rd Qu.: 5324   3rd Qu.: 6.540   3rd Qu.: 6.540  
 Max.   :95.00   Max.   :18823   Max.   :10.740   Max.   :58.900  

       z         
 Min.   : 0.000  
 1st Qu.: 2.910  
 Median : 3.530  
 Mean   : 3.539  
 3rd Qu.: 4.040  
 Max.   :31.800  
```

Graphical Repreentations of Data

- Bar Plot
- Histogram
- Density Plot
- Box Plot
- Scatter Plot

Bar Plot

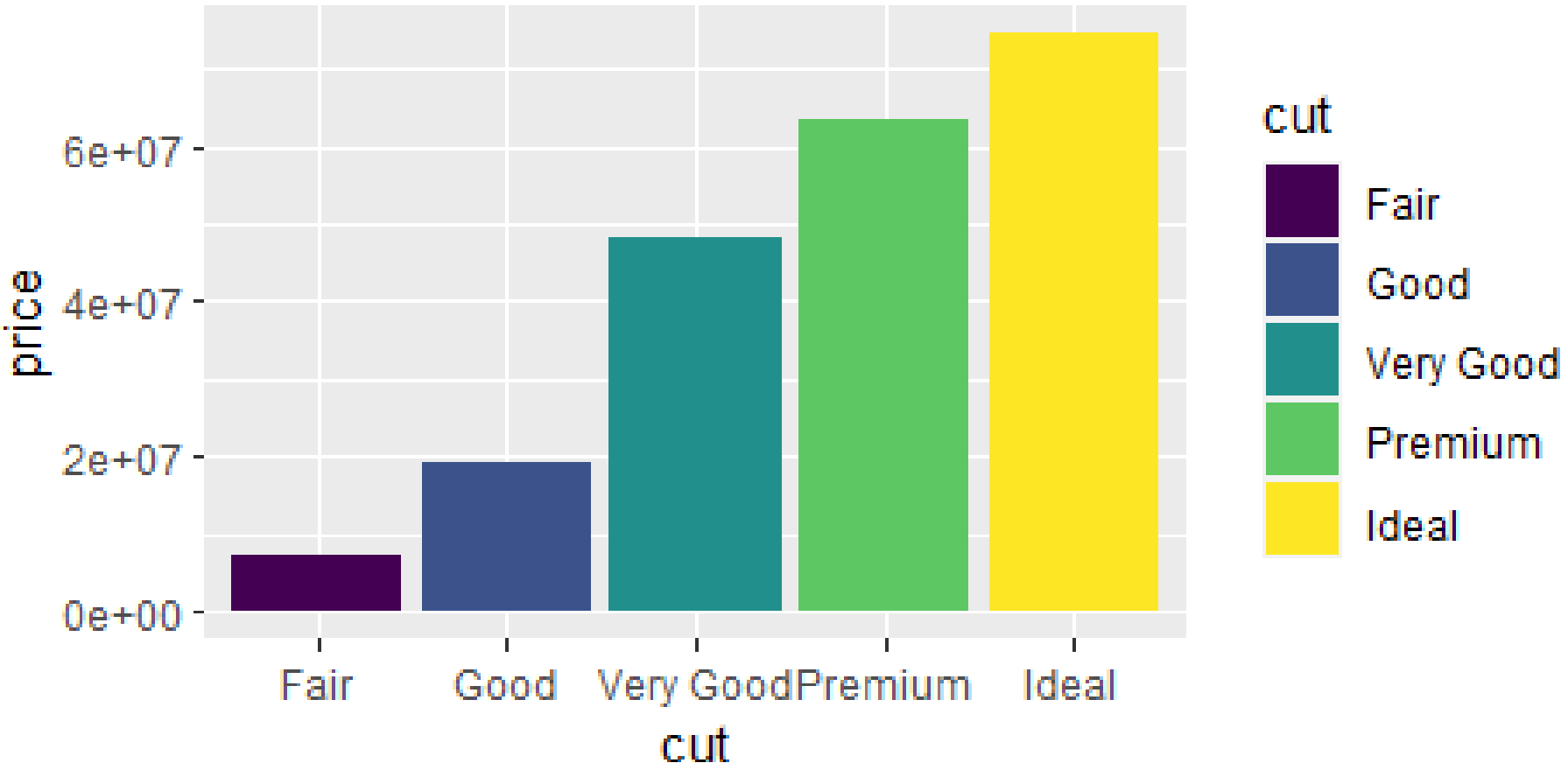


Figure 4: Prices of Various cuts of diamonds

Histogram

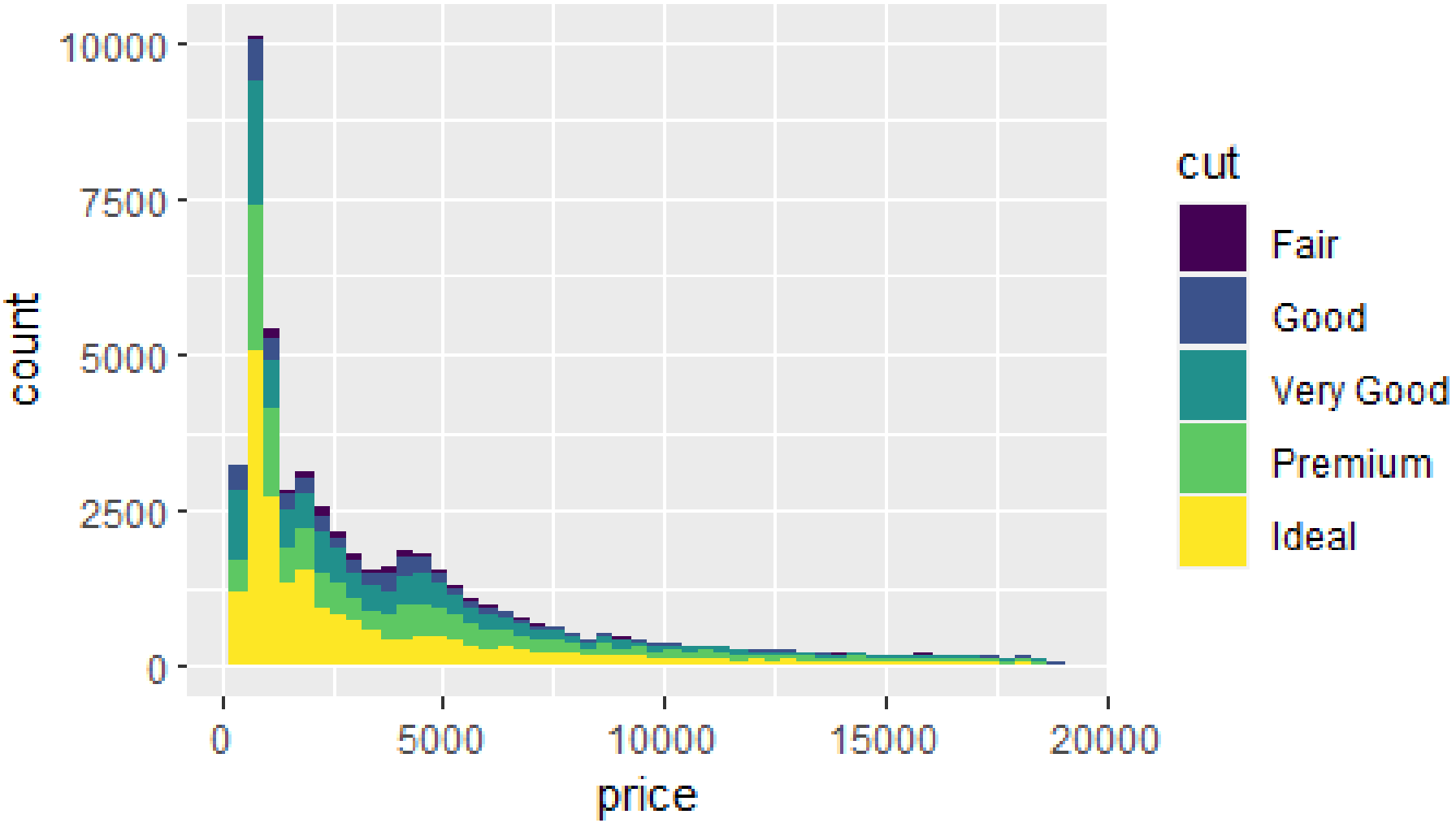


Figure 5: Histogram showing the various cut of diamonds

Density Plot

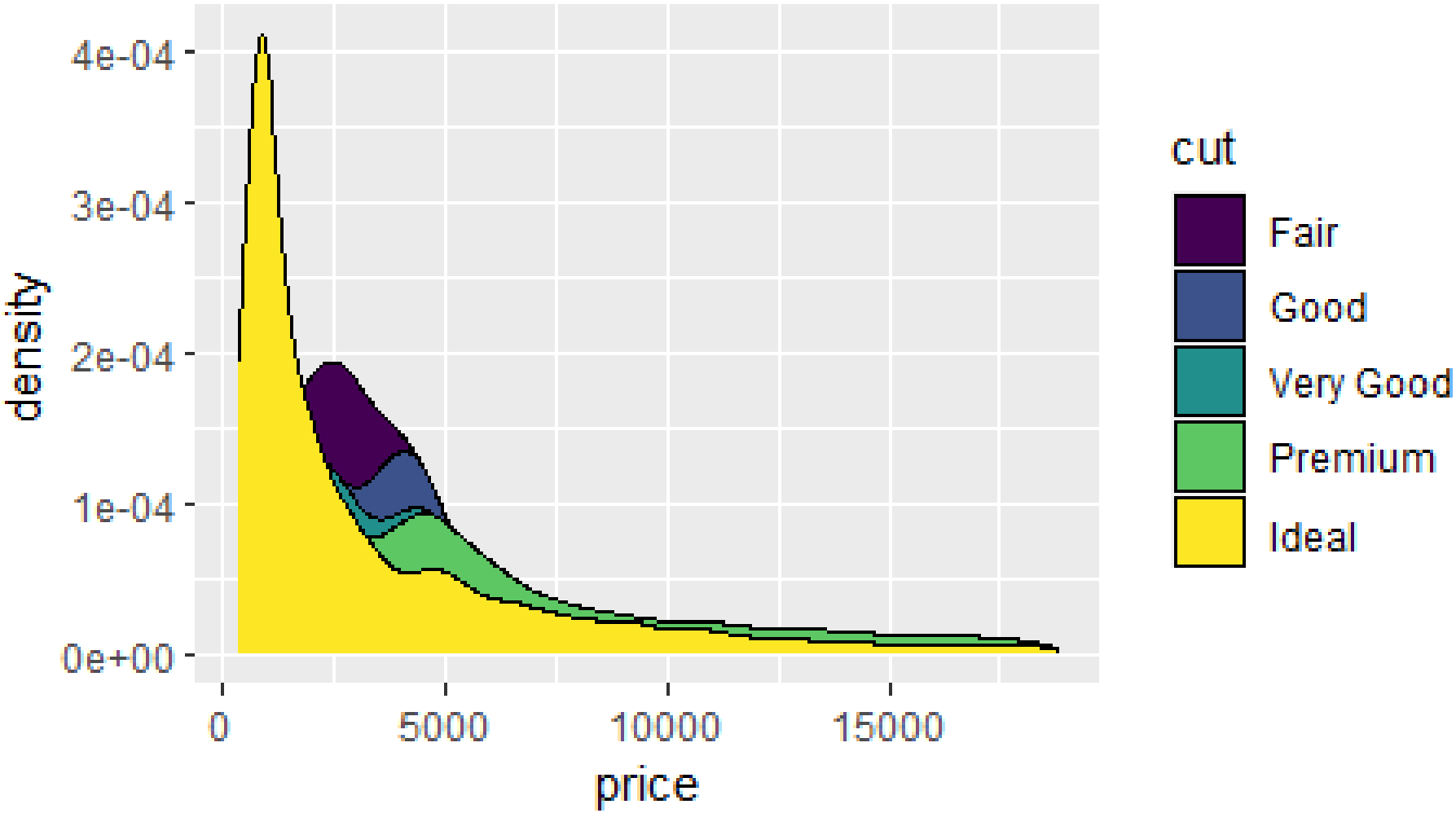


Figure 6: Density Plot of various diamond cut

Box Plot

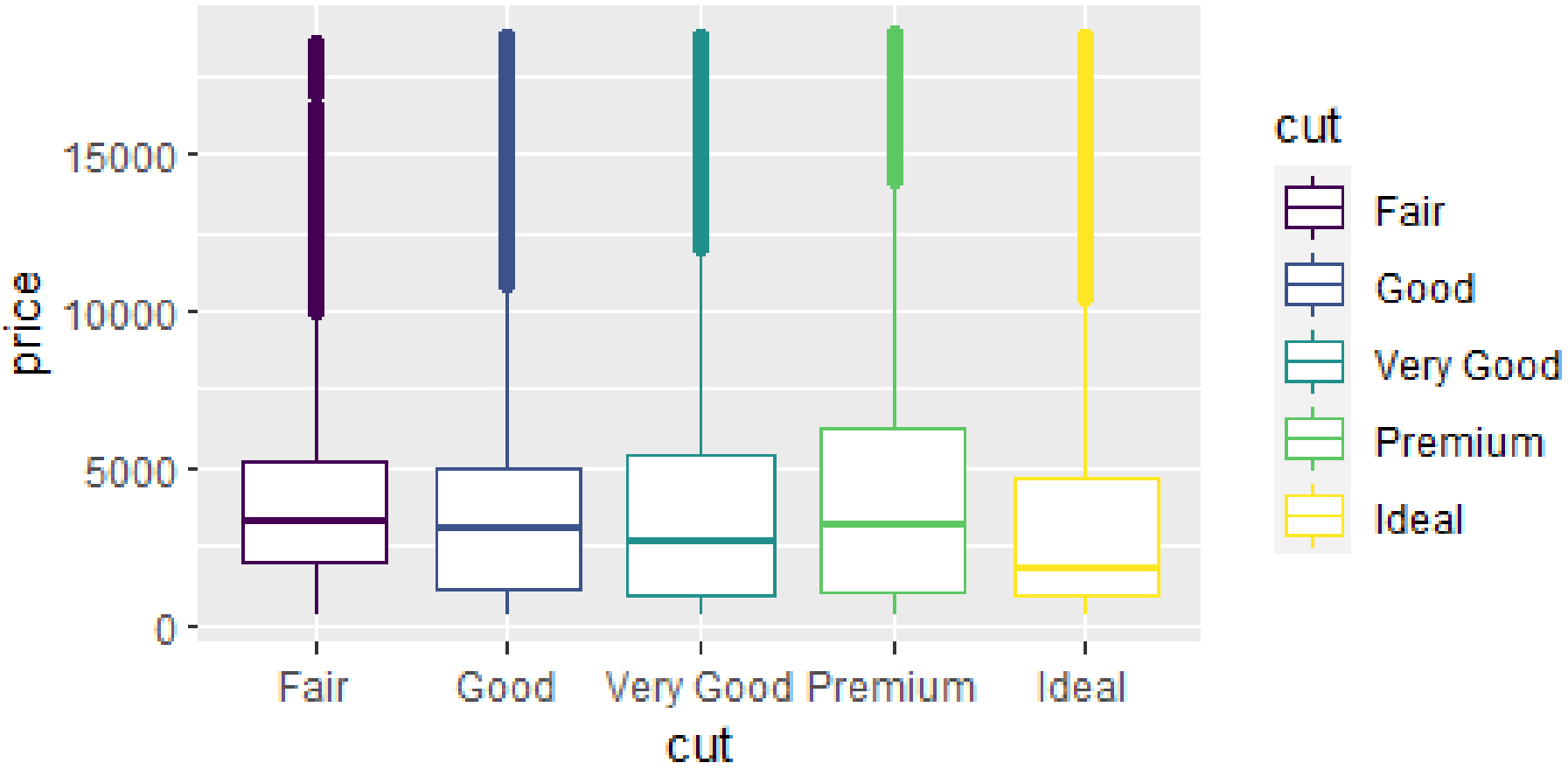


Figure 7: Box plot of various diamond cut

Scatter Plot

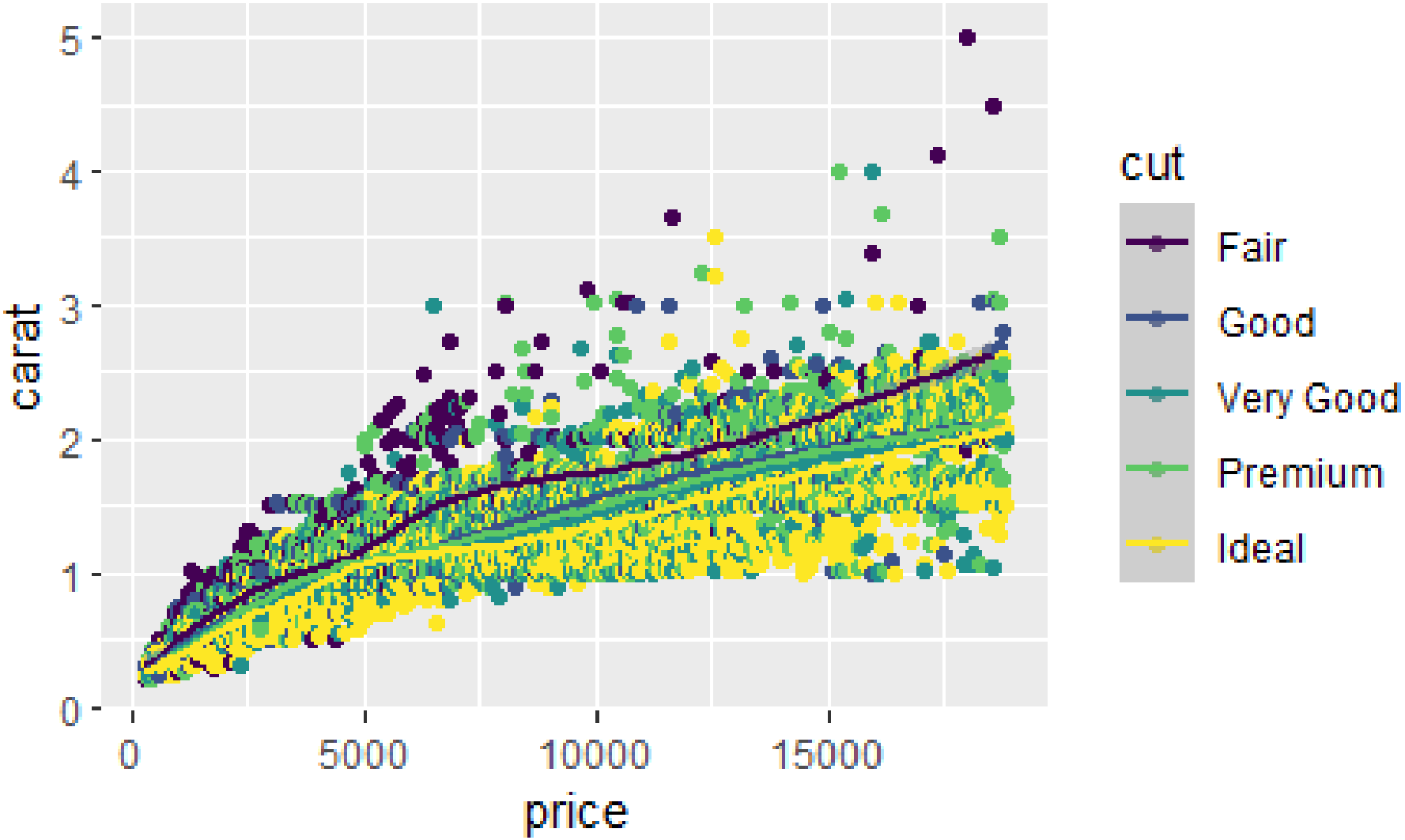


Figure 8: Scatter plot showing the relationship between carat and price

Challenge

Reference