

STOR 320 Exploratory Data Analysis

Lecture 8

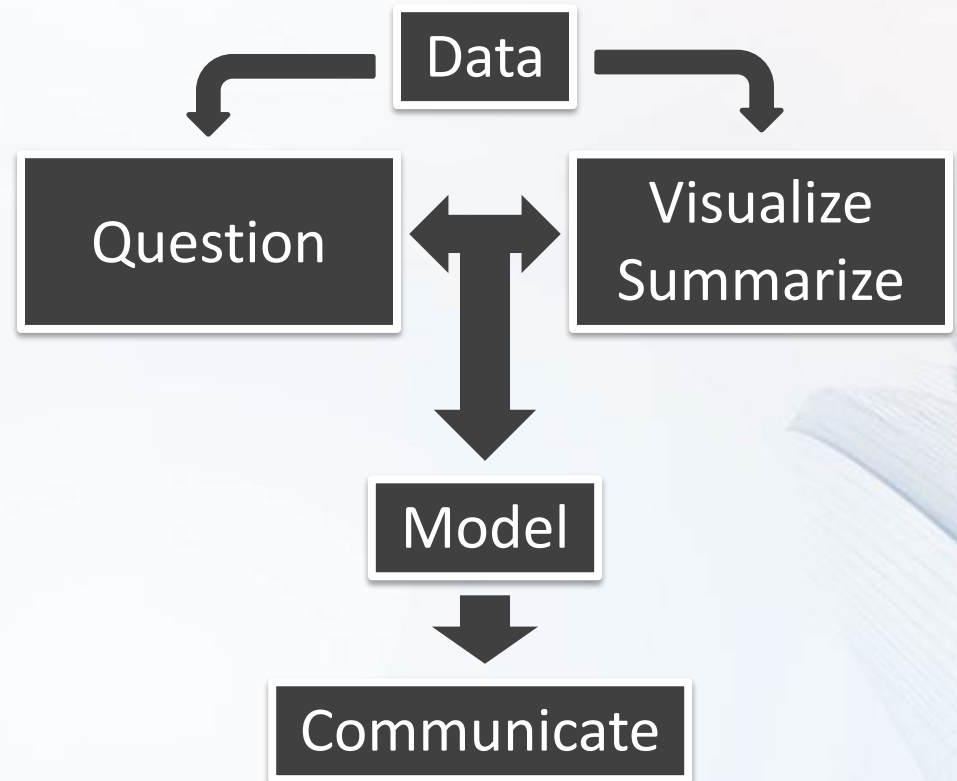
Yao Li

Department of Statistics and Operations Research

UNC Chapel Hill

EDA Definition

- Read Chapter 7
- Know the Process
- Respect the Process



Question

- Think Creatively
- Quantity and Quality
- General:
 - What type of variation occurs **within** my variables?
 - What type of covariation occurs **between** my variables?

Data

- Example: Wages1

- “Ecdat” R Package
- Sample from 1976-1982
 - 3,294 Workers
 - 4 variables
- Variables
 - Experience (Yrs.)
 - Gender (M or F)
 - School (Yrs.)
 - Wage (Hourly in \$)

```
````{r}
wage=as.tibble(wages1) %>%
 rename(experience=exper) %>%
 arrange(school)
head(wage, 10)
````
```

| experience
<int> | gender
<fctr> | school
<int> | wage
<dbl> |
|---------------------|------------------|-----------------|---------------|
| 18 | male | 3 | 5.51682632 |
| 15 | male | 4 | 3.56497766 |
| 18 | male | 4 | 9.09918107 |
| 10 | female | 5 | 0.60316541 |
| 11 | male | 5 | 3.80264284 |
| 14 | male | 5 | 7.50044646 |
| 16 | male | 5 | 4.30366672 |
| 14 | male | 5 | 4.88629309 |
| 15 | female | 6 | 4.30366672 |
| 9 | female | 6 | 2.21160651 |

Verbeek, Marno (2004) A Guide to Modern Econometrics, John Wiley and Sons.

Question

- Variation
 - Variable = Quantity, Quality, or Property You Can Measure
 - Reason: Values Tend to “Vary”
 - Example: Random
 - Categorical:
 - Gender
 - Numerical:
 - Wage
 - Experience
 - School

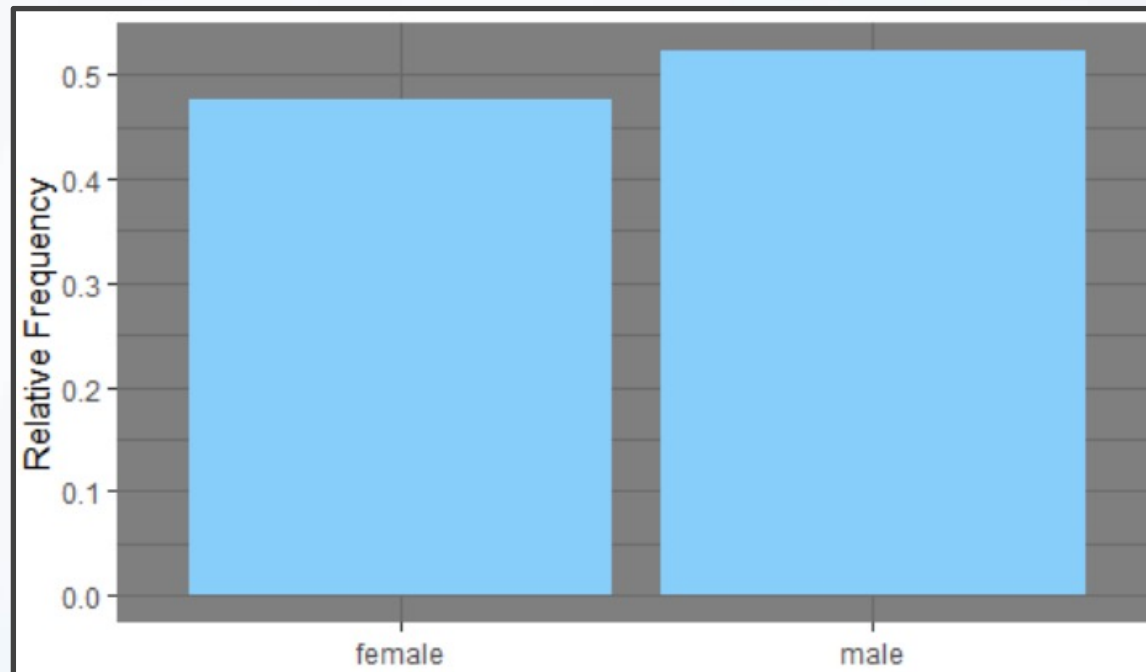
Question

- Initial Questions
 - Example:
 - What did the Workforce Look Like in Terms of Sex?
 - How Spread Out Were Wages?
 - Where is the Middle 50% of the Sample in Regards to Years of Schooling?

Visualize Summarize

- Variation Visualized
- Example: Wages
 - Categorical: Gender

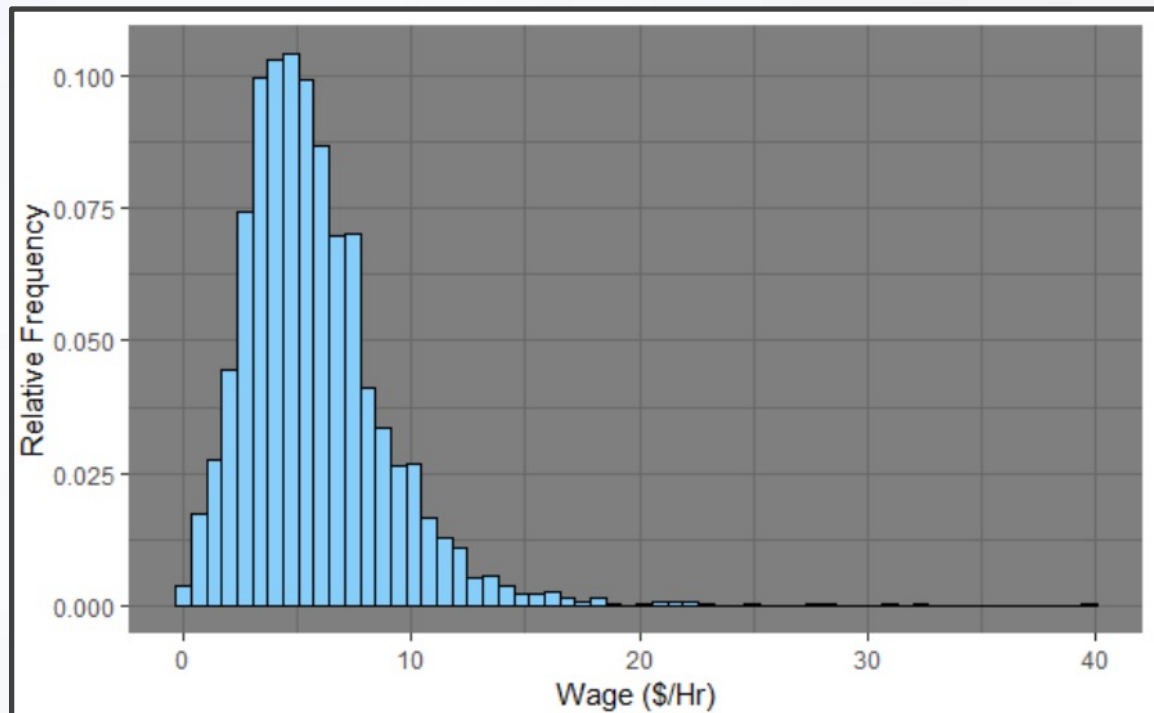
| gender | n |
|--------|-------|
| <fctr> | <int> |
| female | 1569 |
| male | 1725 |



Visualize Summarize

- Variation Visualized
 - Example: Wages
 - Numerical: Hourly Wage

| n
<int> | avg
<dbl> | sd
<dbl> | median
<dbl> | iqr
<dbl> |
|-------------------|---------------------|--------------------|------------------------|---------------------|
| 3294 | 5.757585 | 3.269186 | 5.205781 | 3.682936 |

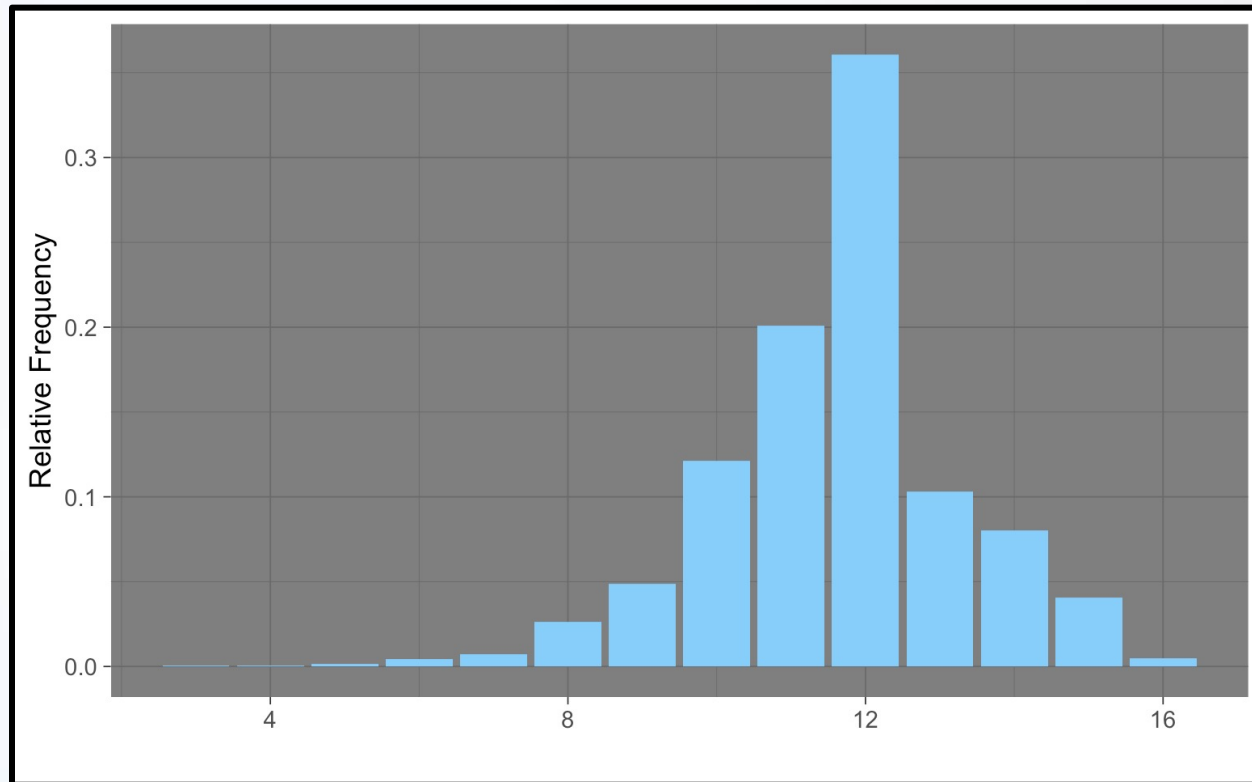


Visualize Summarize

- Variation Visualized

- Example: Wages
 - Numerical: School

| n | avg | sd | median | q1 | q3 | iqr |
|----------|------------|-----------|---------------|-----------|-----------|------------|
| <int> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> |
| 3294 | 11.63054 | 1.657545 | 12 | 11 | 12 | 1 |



Unusual Values

- Outliers = Observations Outside the Pattern of the Data
- Due to Error ➡ Remove
- Don't Drop or Change Without Justification
- Sensitivity Analysis
- Handling:
 - Drop Entire Row
 - Replace Instance with NA

Unusual Values

- Example: Wages
 - Few People Above 30 \$/Hr

- Drop Entire Row

```
```{r}
wage2=wage %>%
 filter(between(wage,0,30))
```

Observations: 3294 ➡ 3291

- Replace Instance with NA

```
```{r}
wage3=wage %>%
  mutate(wage=ifelse(wage>30,NA,wage))
```

Observations: 3294 ➡ 3294

Question

- Covariation
 - Goal: Explain Covariation
 - Describes the Behavior Between Variables
 - We Often Attempt to Explain Variation **Within** by Looking at Covariation **Between**
 - Identify the **Signal** despite the **Noise**

Data



- Example: diamonds
 - “ggplot2” R Package
 - Sample from 1976-1982
 - 53, 940 diamonds
 - 10 variables
- Variables
 - carat
 - cut
 - color
 - clarity
 - depth
 - table
 - price
 - x, y, z

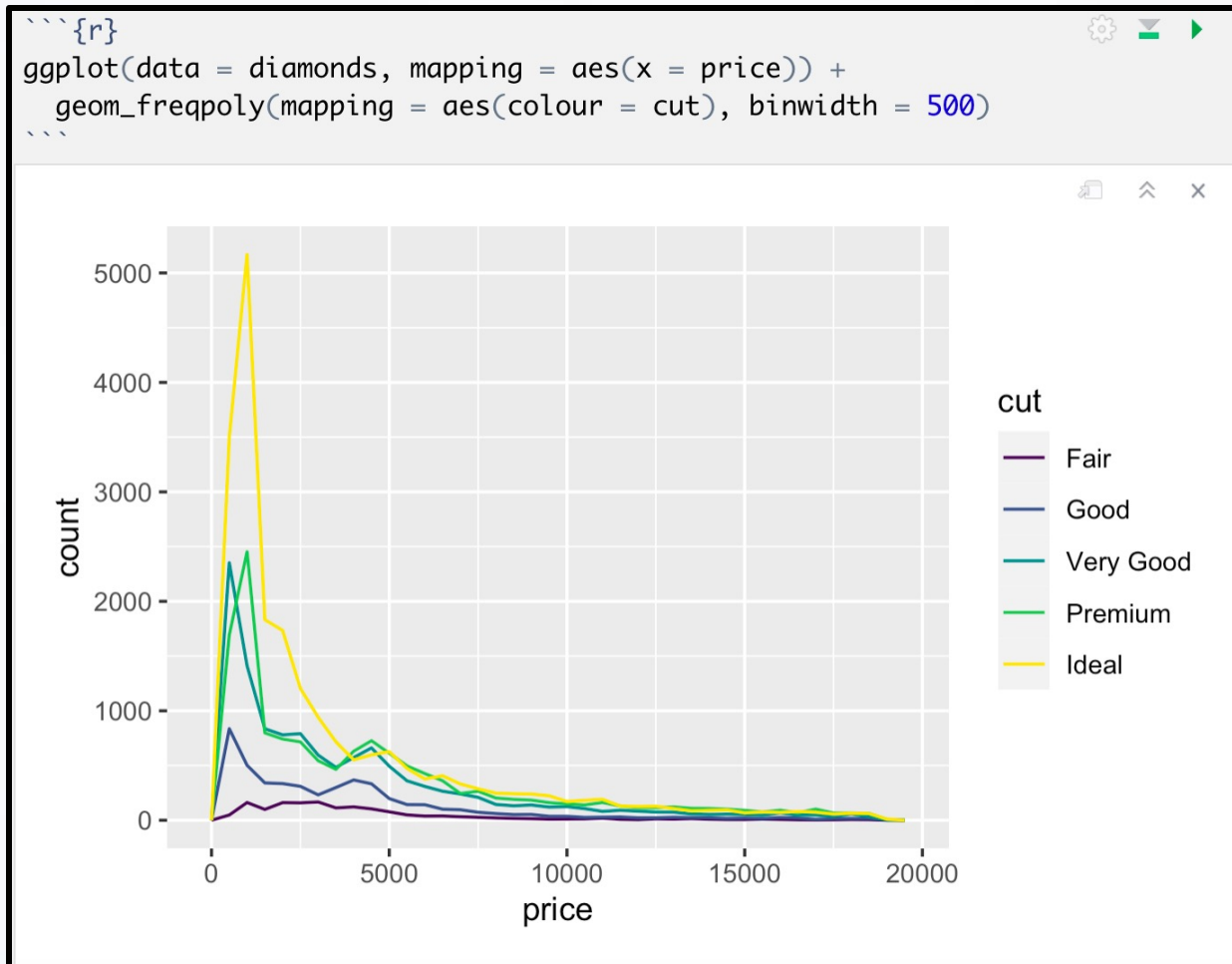
carat <dbl>	cut <ord>	color <ord>	clarity <ord>	depth <dbl>	table <dbl>	price <int>	x <dbl>	y <dbl>	z <dbl>
0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75
0.24	Very Good	J	VVS2	62.8	57.0	336	3.94	3.96	2.48
0.24	Very Good	I	VVS1	62.3	57.0	336	3.95	3.98	2.47
0.26	Very Good	H	SI1	61.9	55.0	337	4.07	4.11	2.53
0.22	Fair	E	VS2	65.1	61.0	337	3.87	3.78	2.49
0.23	Very Good	H	VS1	59.4	61.0	338	4.00	4.05	2.39

Question

- Covariation Questions
 - Example: Wages
 - Does Quality of a diamond affect Price?
 - Does Color Affect Quality?
 - What is the Relationship Between Weight and Price?

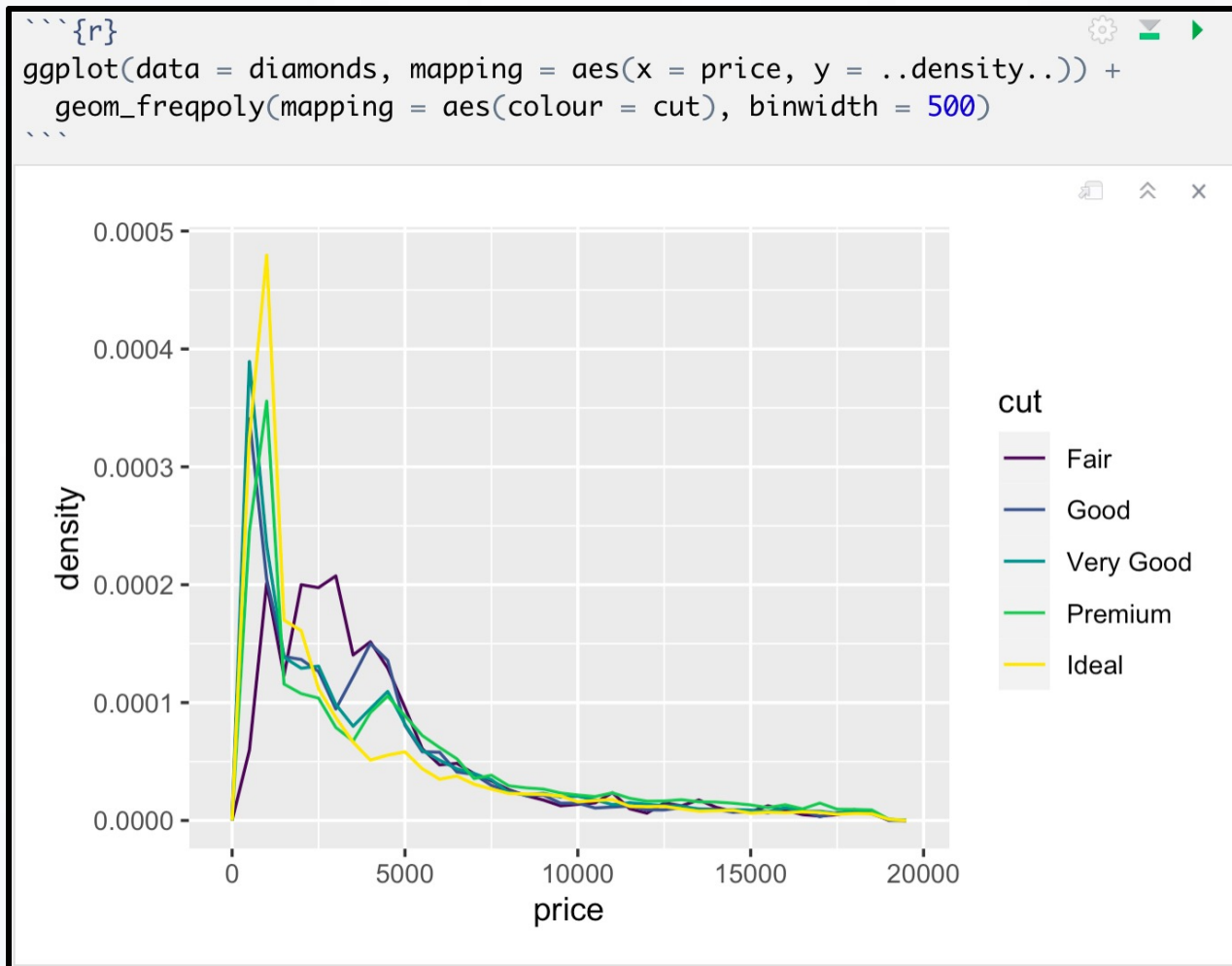
Visualize Summarize

- Categorical and Continuous



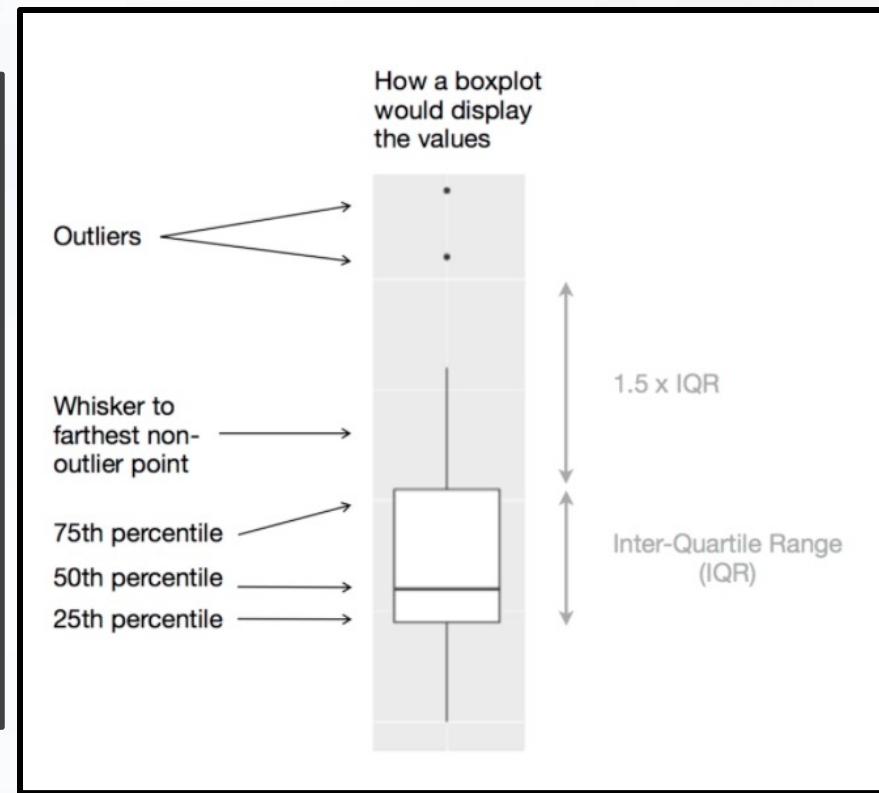
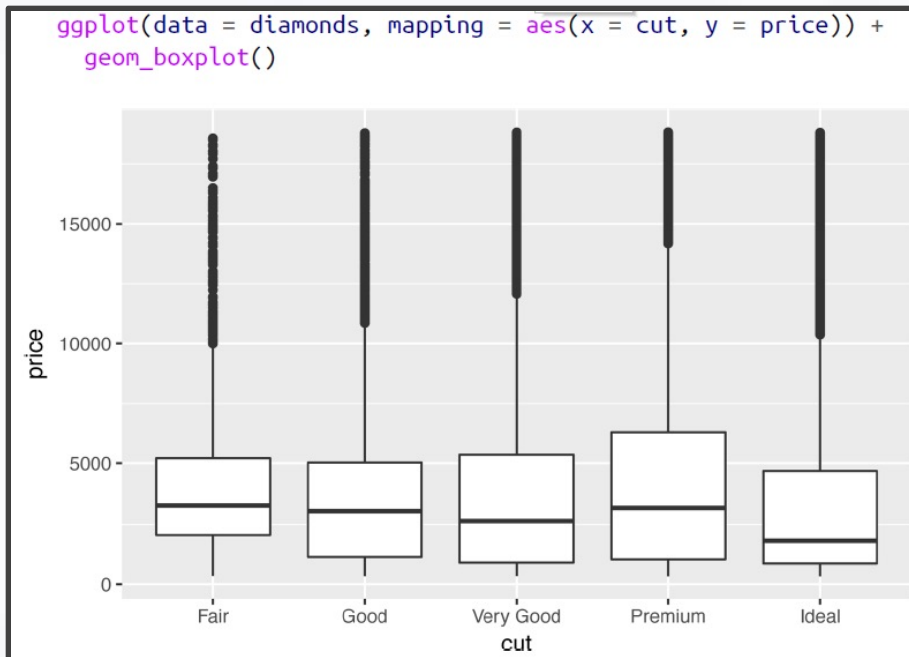
Visualize Summarize

- Categorical and Continuous: density



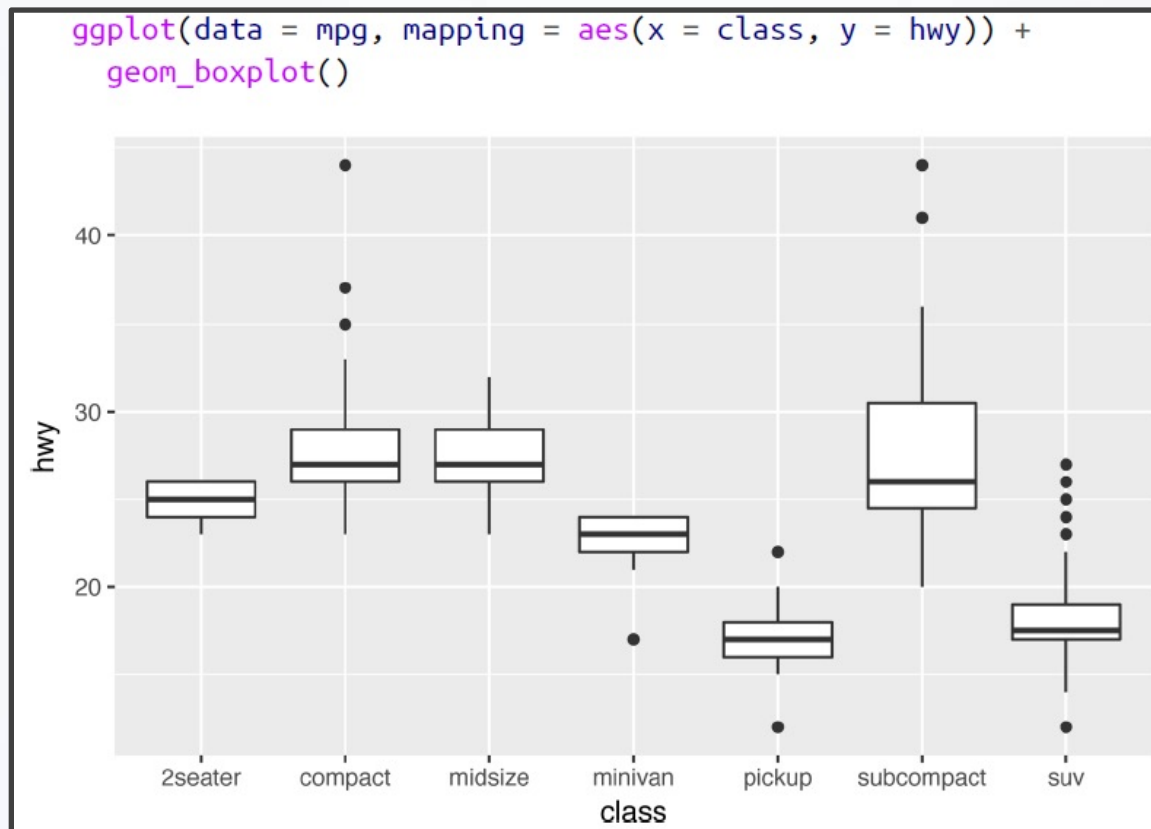
Visualize Summarize

- Categorical and Continuous



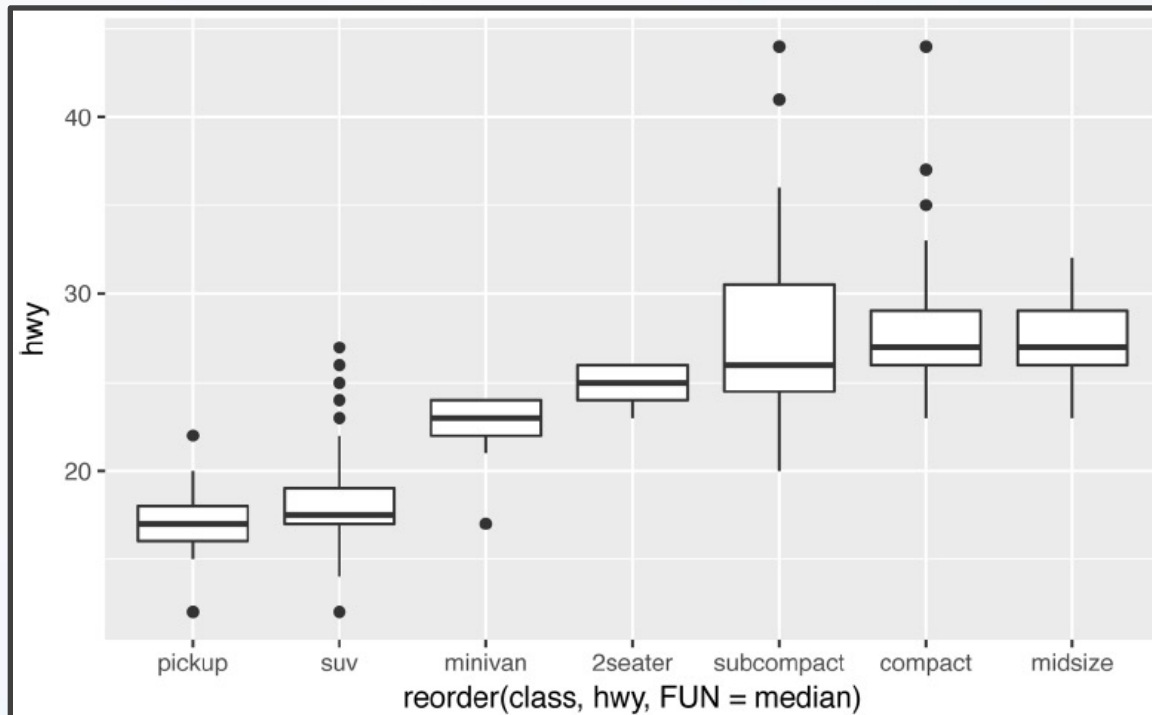
Visualize Summarize

- Categorical and Continuous



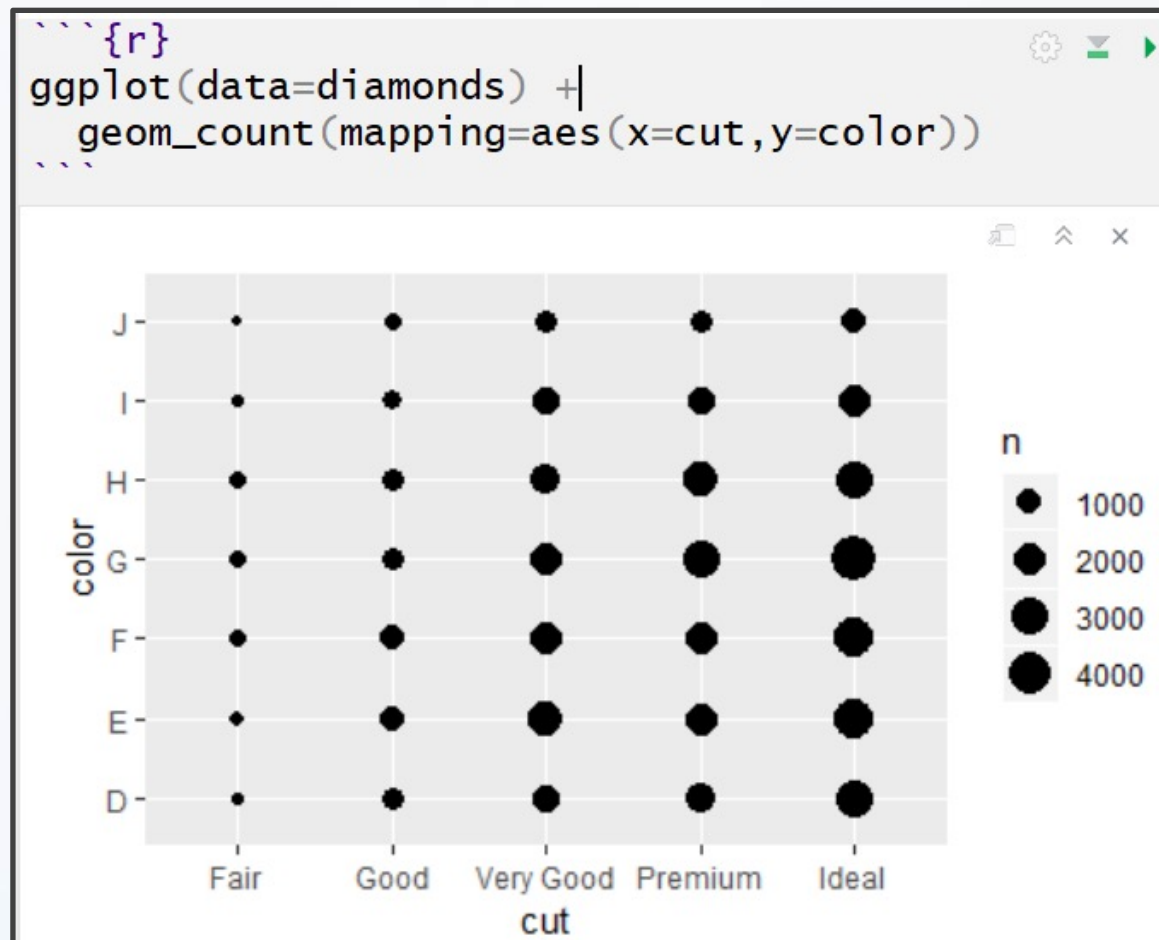
- Categorical and Continuous

```
ggplot(data = mpg) +  
  geom_boxplot(  
    mapping = aes(  
      x = reorder(class, hwy, FUN = median),  
      y = hwy  
    )  
  )  
)
```



Visualize Summarize

- Categorical and Categorical



Visualize Summarize

- Categorical and Categorical

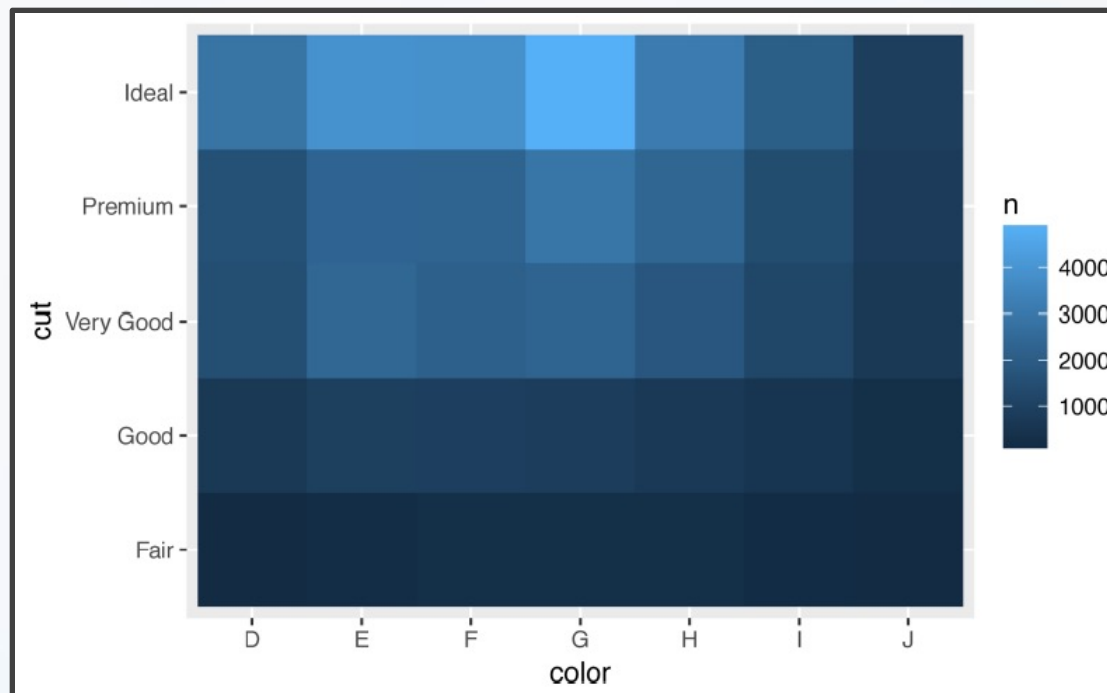
```
```{r}
diamonds %>%
 group_by(cut, color) %>%
 summarize(n=n()) %>%
 spread(cut, n)
```
```

| color
<ord> | Fair
<int> | Good
<int> | Very Good
<int> | Premium
<int> | Ideal
<int> |
|----------------|---------------|---------------|--------------------|------------------|----------------|
| D | 163 | 662 | 1513 | 1603 | 2834 |
| E | 224 | 933 | 2400 | 2337 | 3903 |
| F | 312 | 909 | 2164 | 2331 | 3826 |
| G | 314 | 871 | 2299 | 2924 | 4884 |
| H | 303 | 702 | 1824 | 2360 | 3115 |
| I | 175 | 522 | 1204 | 1428 | 2093 |
| J | 119 | 307 | 678 | 808 | 896 |

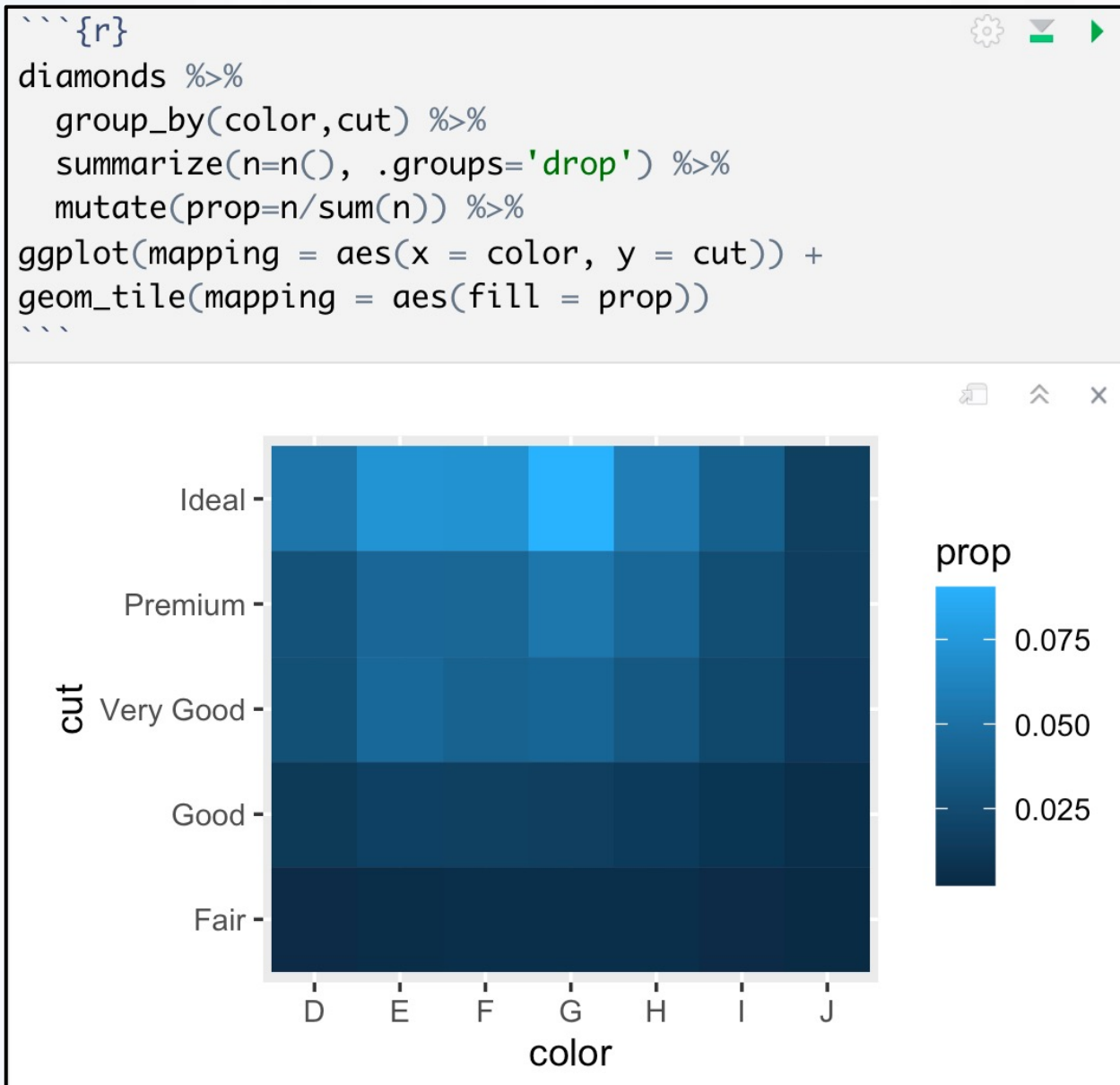
Visualize Summarize

- Categorical and Categorical

```
diamonds %>%  
  count(color, cut) %>%  
  ggplot(mapping = aes(x = color, y = cut)) +  
    geom_tile(mapping = aes(fill = n))
```



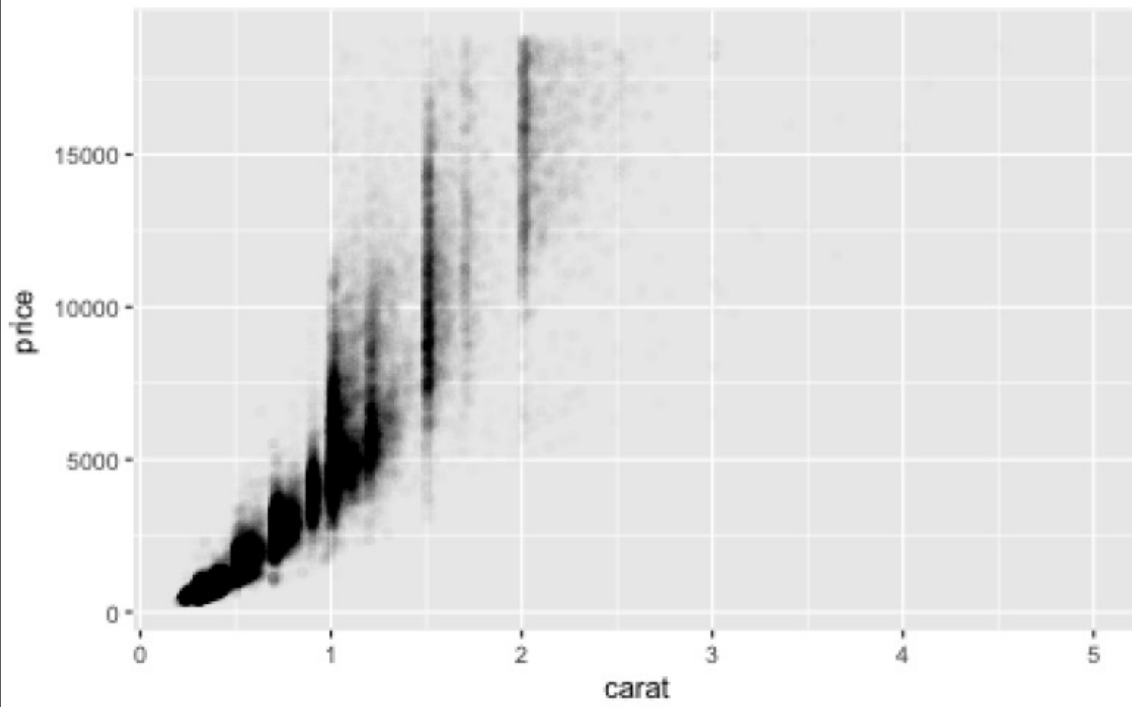
- Categorical and Categorical



Visualize Summarize

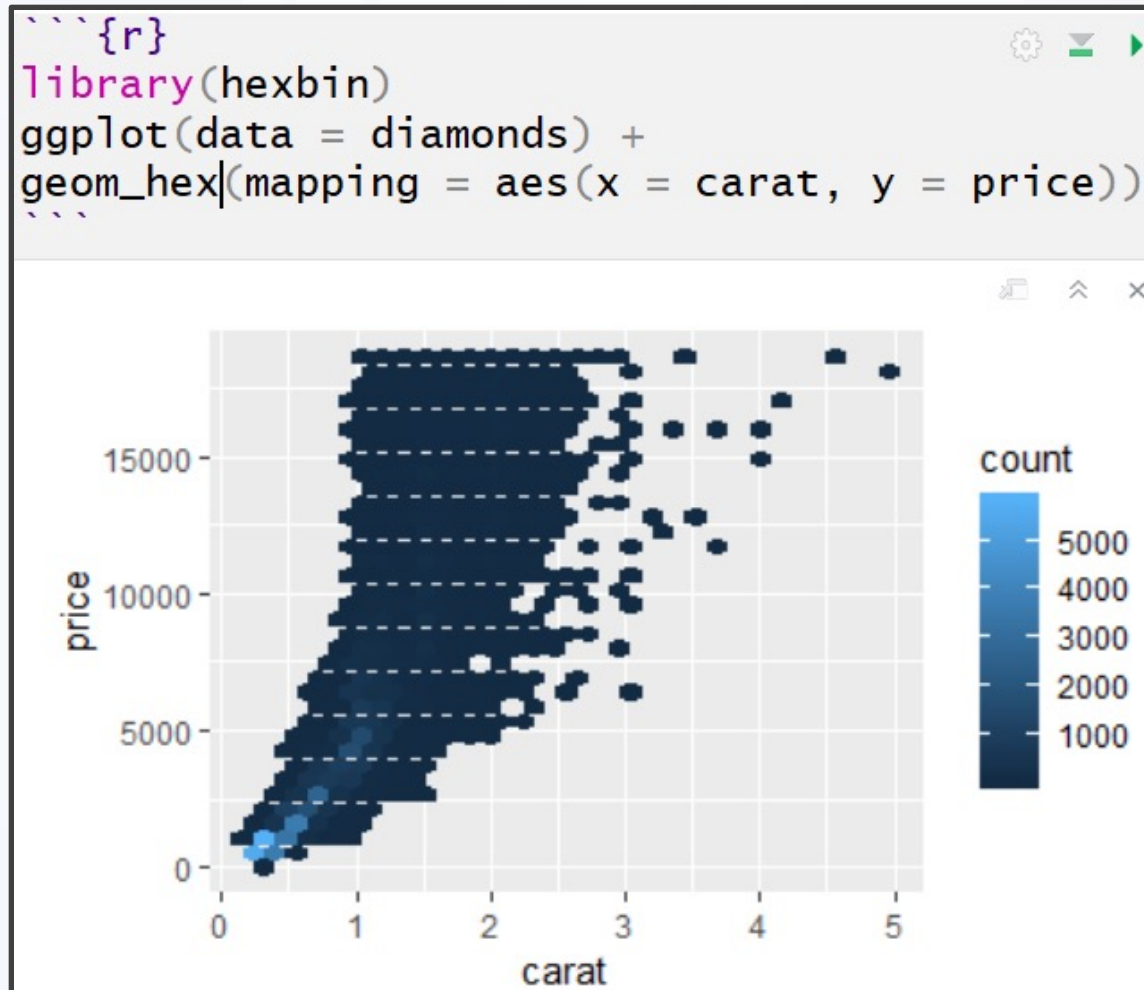
- Continuous and Continuous

```
ggplot(data = diamonds) +  
  geom_point(  
    mapping = aes(x = carat, y = price),  
    alpha = 1 / 100  
  )
```



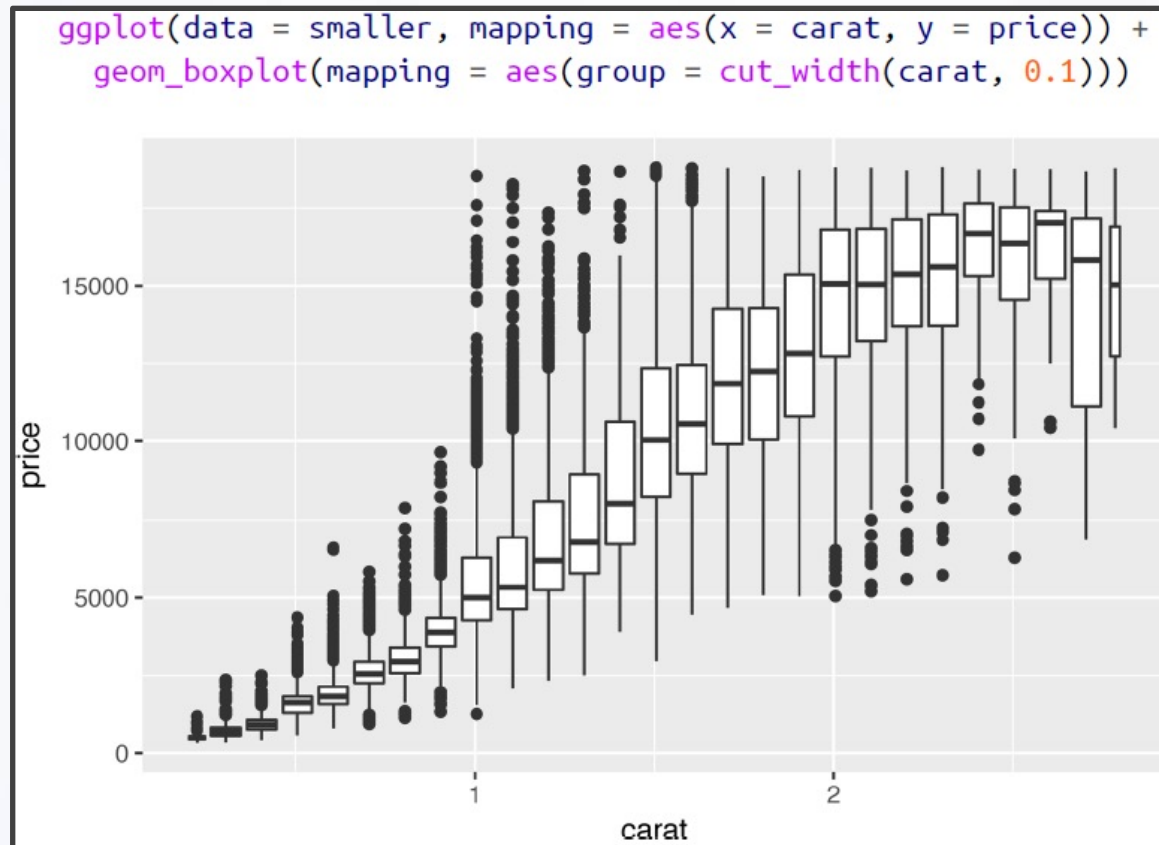
Visualize Summarize

- Continuous and Continuous



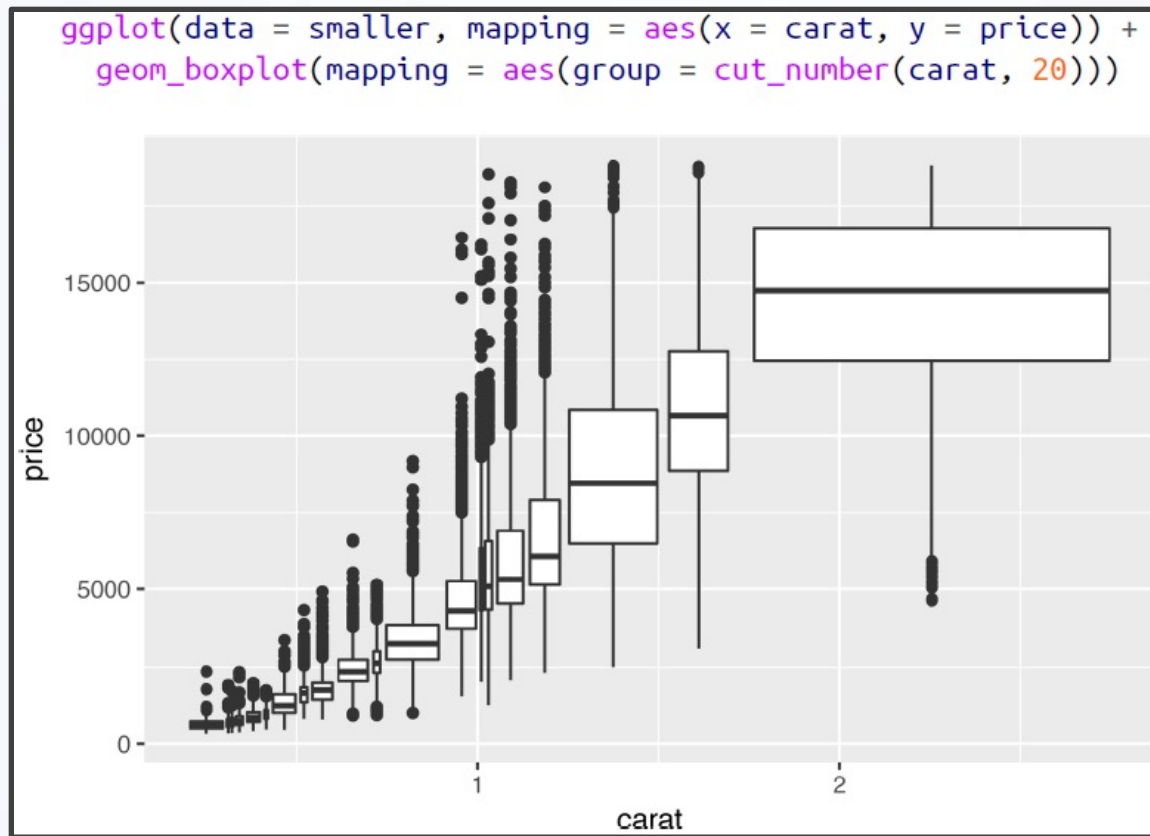
Visualize Summarize

- Continuous and Continuous



Visualize Summarize

- Continuous and Continuous

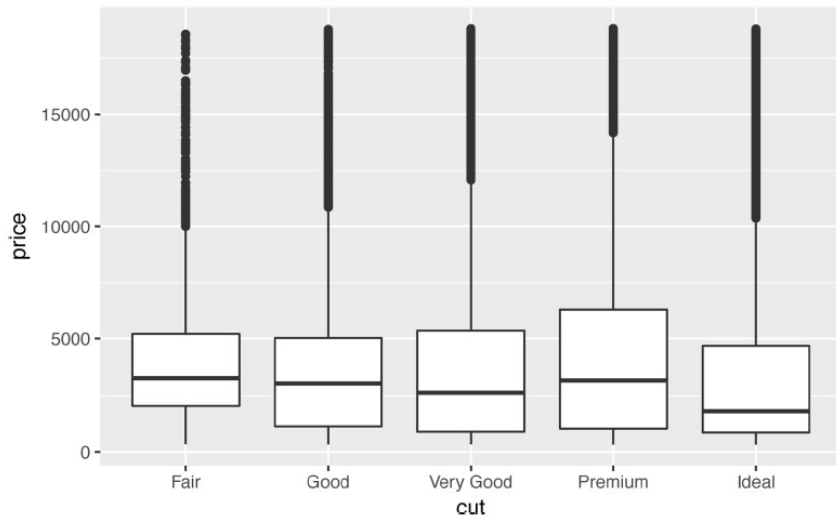


EDA Purpose

- Purpose of Asking Questions and Exploring Those Questions Using Visualizations and Summaries is to Spot Patterns
- Ask Yourself:
 - Is it Coincidence?
 - How Strong is the Relationship?
 - What Variables May Be Confounding?
 - Do Subgroups Cause the Relationship to Change?
 - How Can You Model the Pattern?

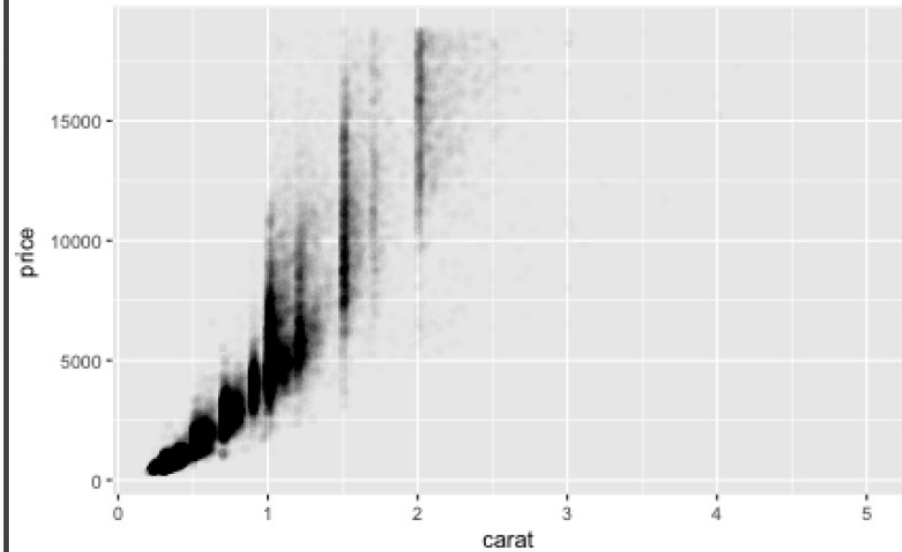
Findings

```
ggplot(data = diamonds, mapping = aes(x = cut, y = price)) +  
  geom_boxplot()
```



- Negative relationship between cut and price

```
ggplot(data = diamonds) +  
  geom_point(  
    mapping = aes(x = carat, y = price),  
    alpha = 1 / 100  
  )
```



- Positive relationship between size and price

Question

What is the relationship between

the size of the



and

the price of the



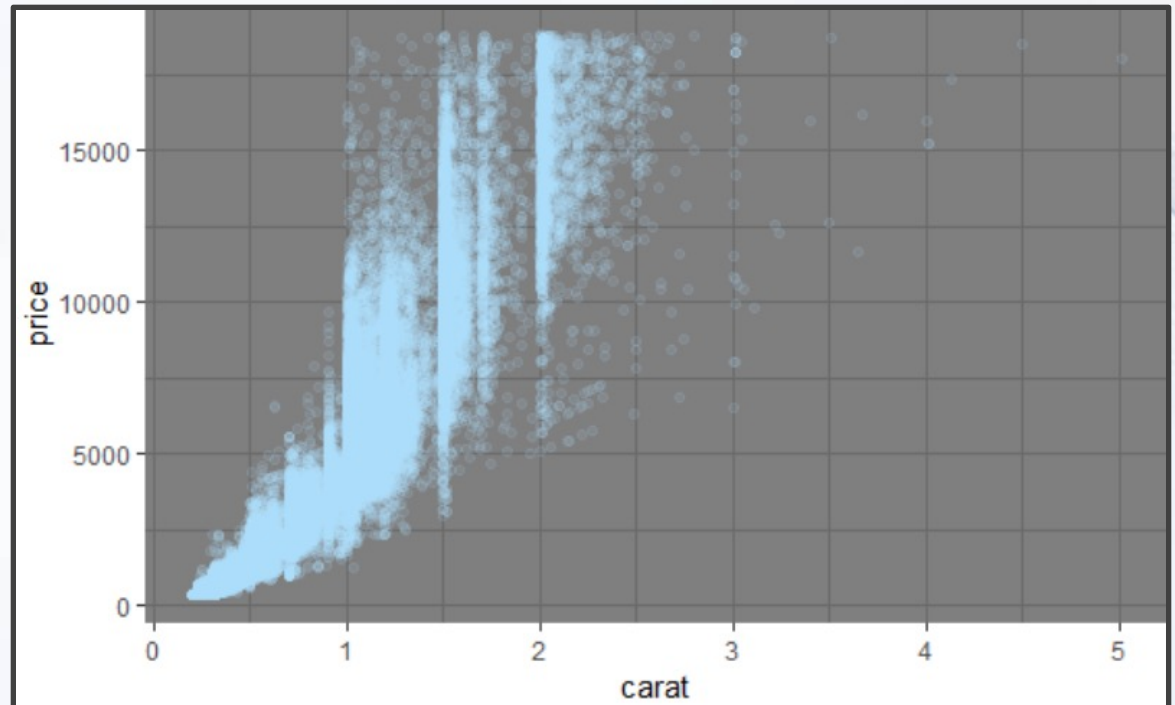
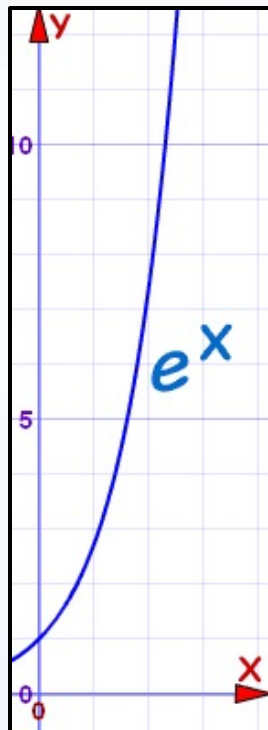
?

Visualize Summarize

```
```{r}
diamonds %>%
 summarize(n=n(),avgprice=mean(price),sdprice=sd(price),
 avgcarat=mean(carat),sdcarat=sd(carat),
 correlation=cor(price,carat))|
...

```

n	avgprice	sdprice	avgcarat	sdcarat	correlation
<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
53940	3932.8	3989.44	0.7979397	0.4740112	0.9215913



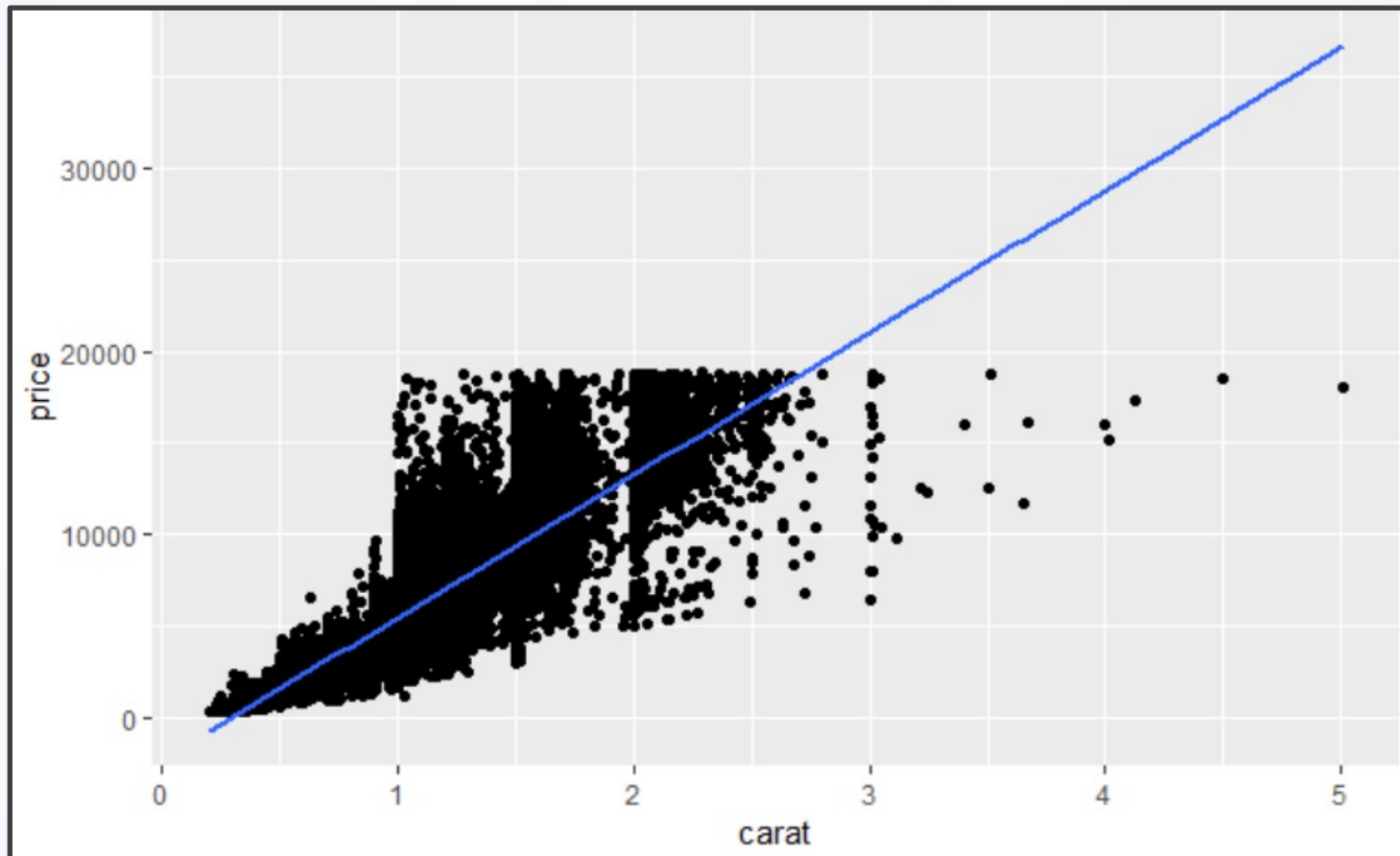
# Question

- Refined Questions
  - Is the Observed Relationship Spurious?
  - Can I Represent the Relationship Using a Linear Model?
  - Should I Use an Exponential Model to Represent the Relationship?
  - Does Another Variable Exist to Explain the Drastic Change in Spread?



# Model

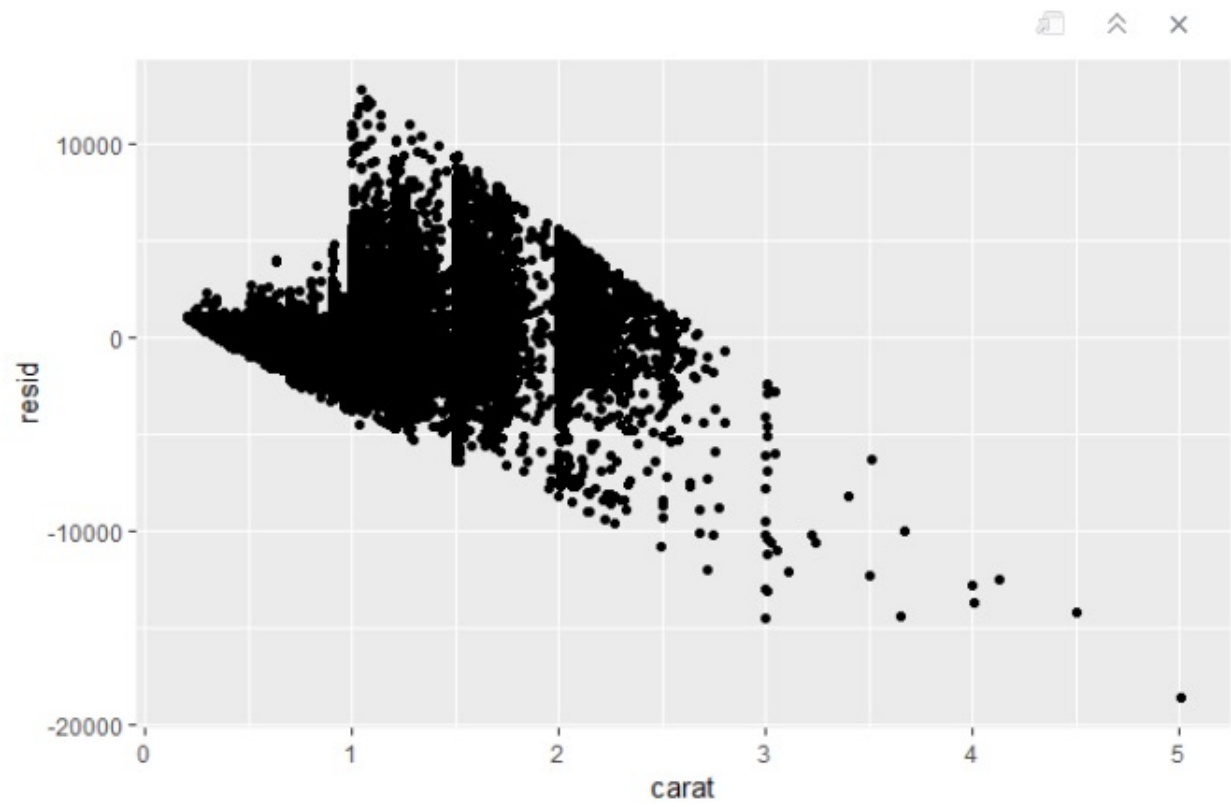
- Linear Model



# Model

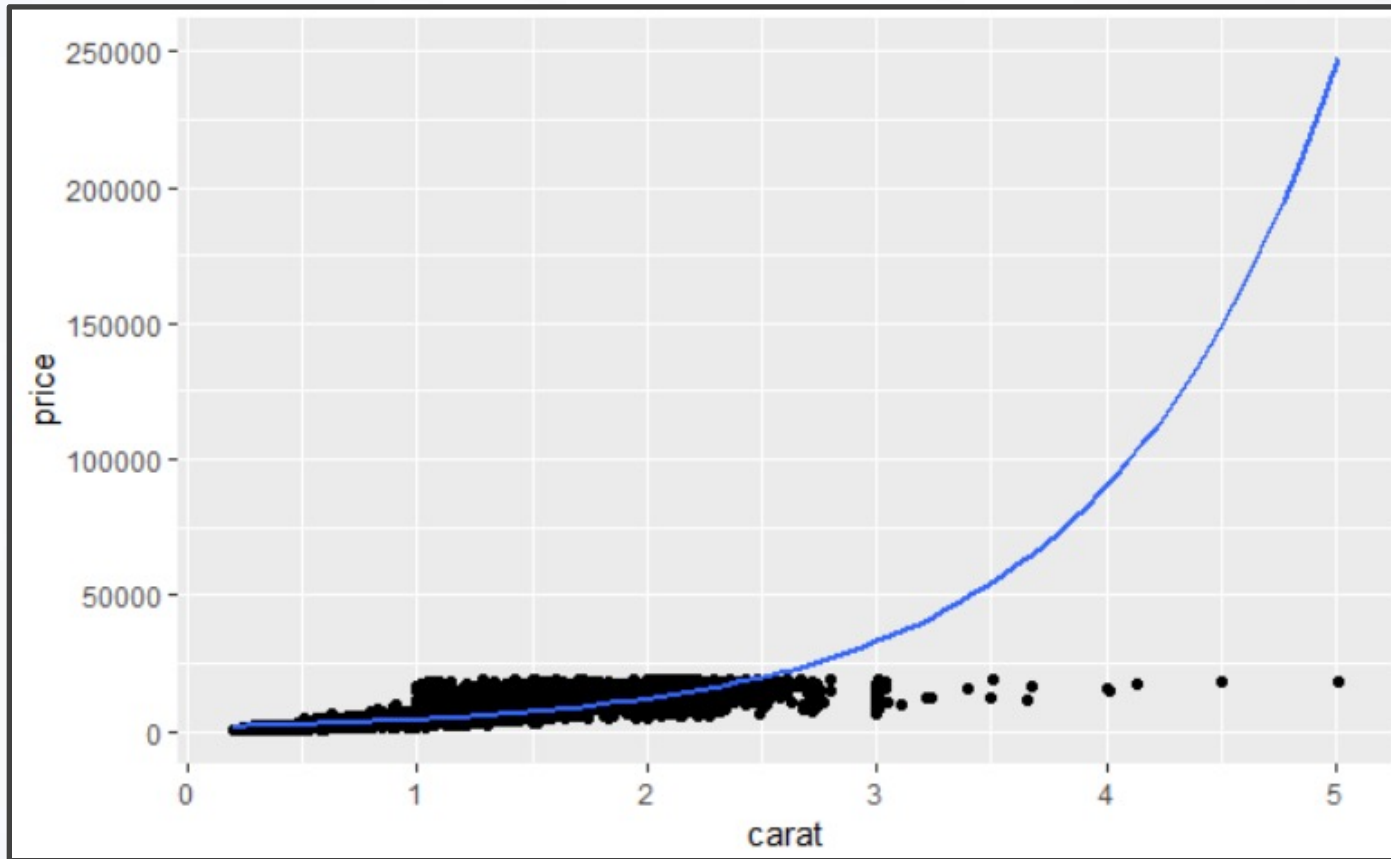
- Linear Model

```
```\r\nlibrary(modelr)\nlin.mod=lm(price~carat,data=diamonds)\ndiamonds.lin.resid = diamonds %>%\n  add_residuals(mod=lin.mod)\nggplot(data=diamonds.lin.resid) +\n  geom_point(aes(x=carat,y=resid))\n```\n
```



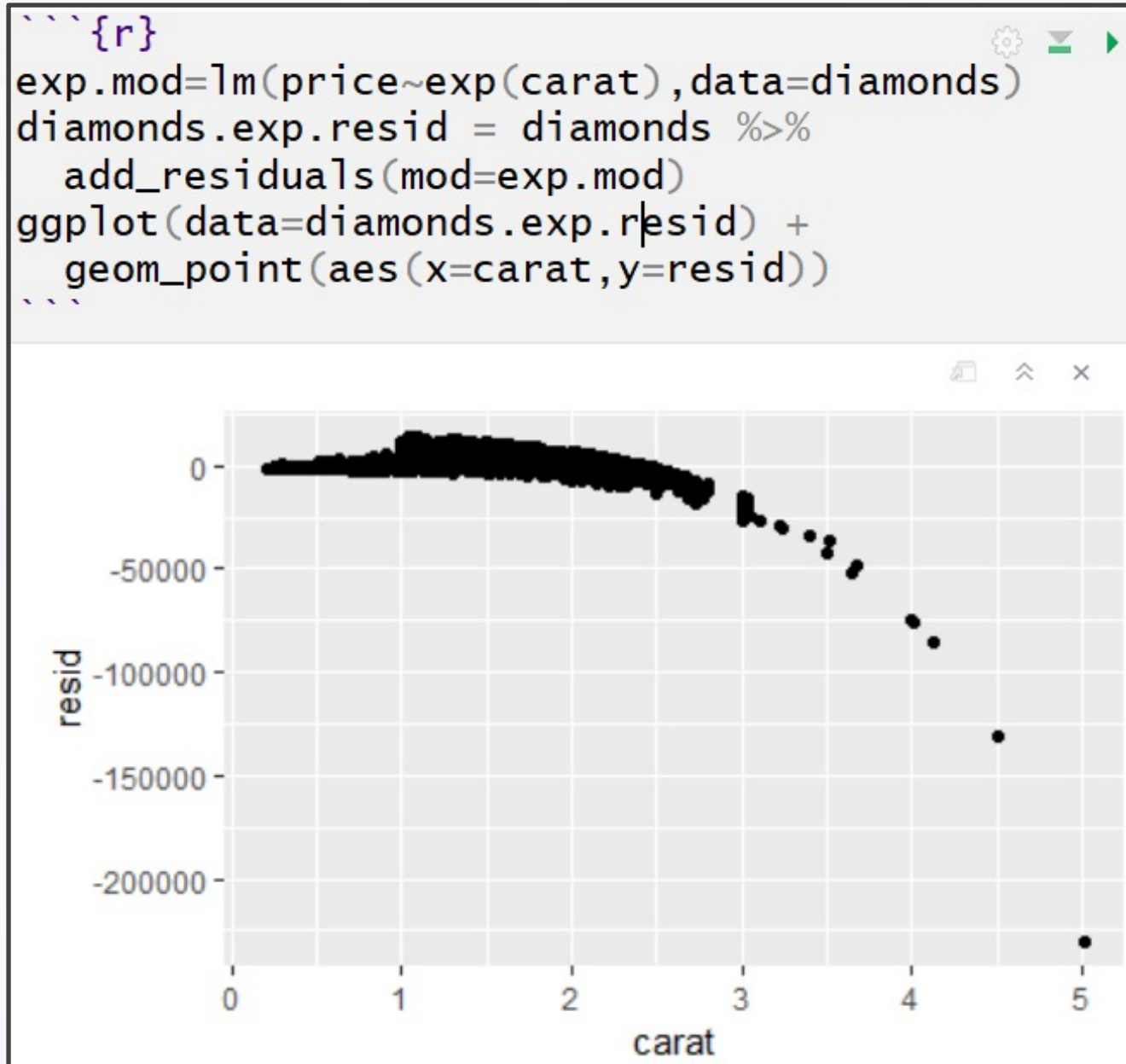
Model

- Exponential Model



Model

- Exponential Model



Model

- Exponential Model

```
```{r}
exp.mod=lm(price~exp(carat),data=diamonds)
diamonds.exp.resid = diamonds %>%
 add_residuals(mod=exp.mod)
ggplot(data=diamonds.exp.resid) +
 geom_point(aes(x=carat,y=resid)) +
 coord_cartesian(xlim=c(0,2.5),
 ylim=c(-25000,25000))
```
```

