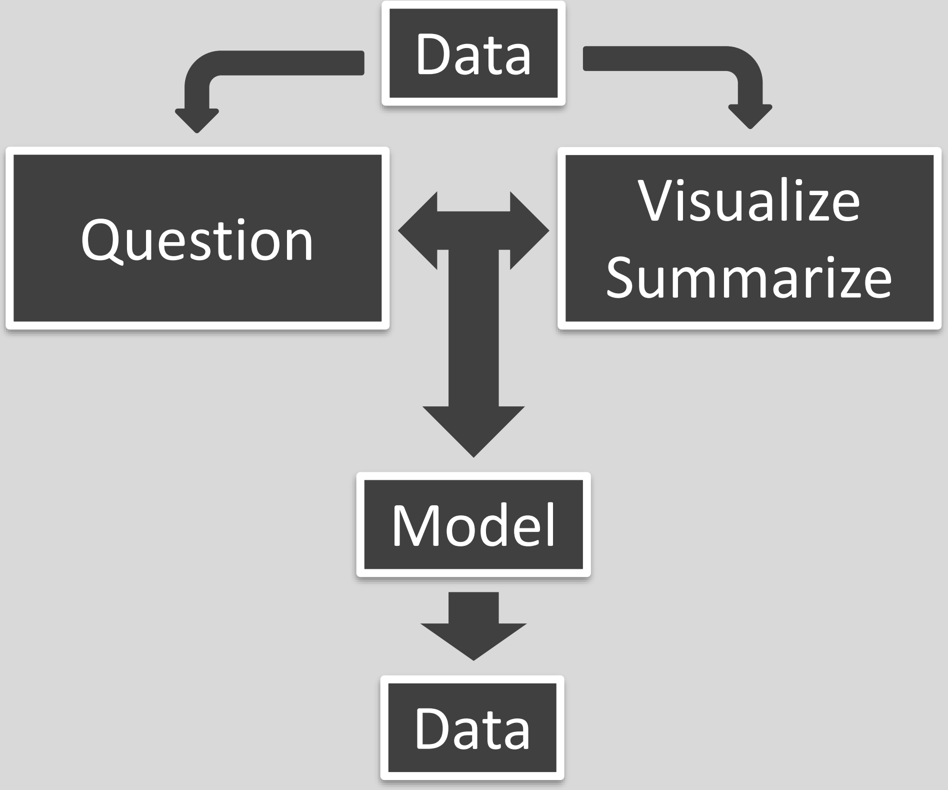
Exploratory Data Analysis I

EDA Defined

- Tenderly Read Chapter 5
- Know the Process

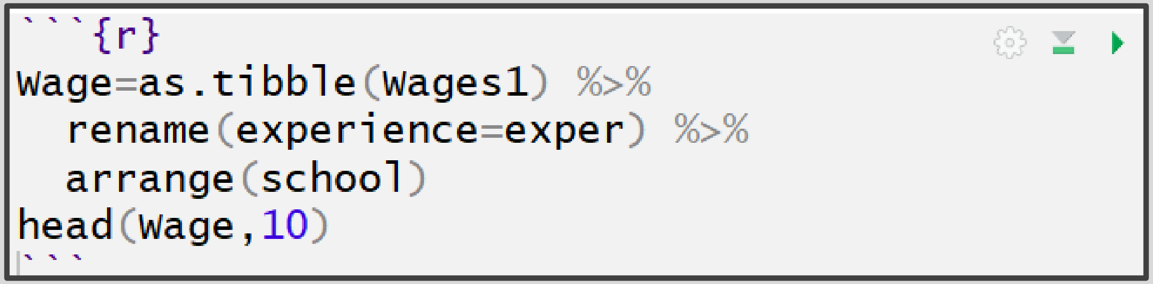


- Respect the Process

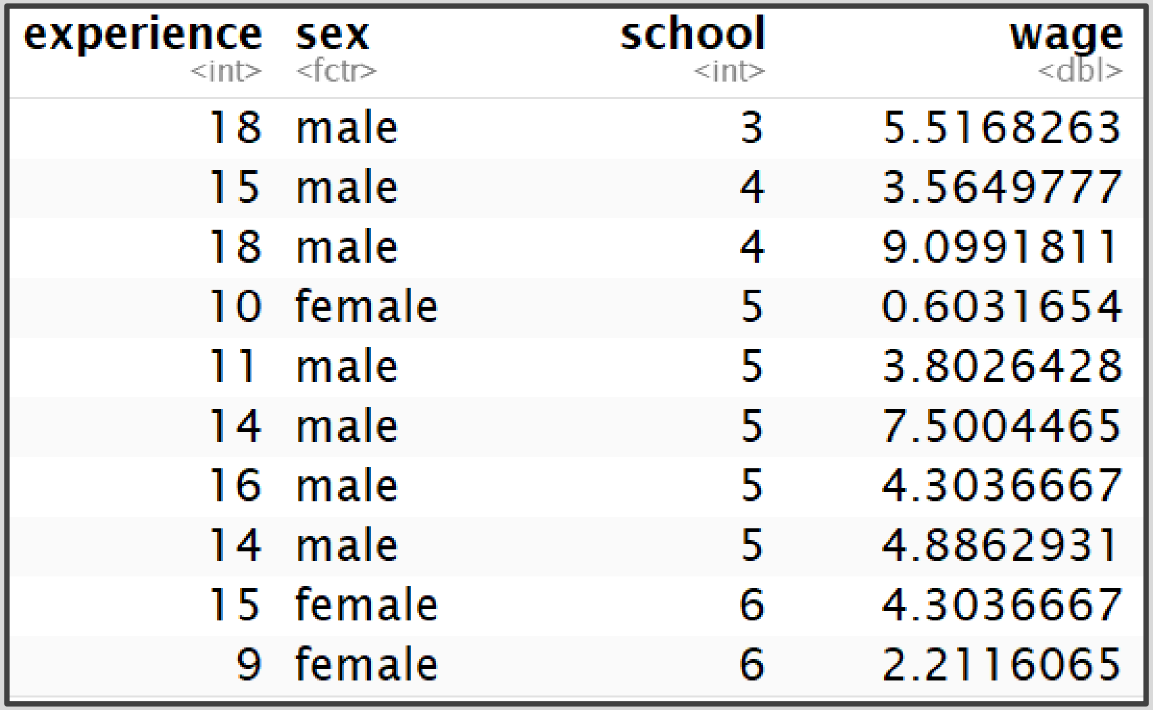
Data

- Example: Wages
  - "Ecdat" R Package
  - Sample from 1987
    - 3,294 Workers
    - 48% Female
  - Variables
    - Experience (Yrs.)
    - Sex (M or F)
    - School (Yrs.)
    - Wage (Hourly in $)

Data



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



```
experience sex    school      wage
     <int> <fctr>  <int>     <dbl>
        18 male        3 5.5168263
        15 male        4 3.5649777
        18 male        4 9.0991811
        10 female      5 0.6031654
        11 male        5 3.8026428
        14 male        5 7.5004465
        16 male        5 4.3036667
        14 male        5 4.8862931
        15 female      6 4.3036667
         9 female      6 2.2116065
```

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

Question

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

Question

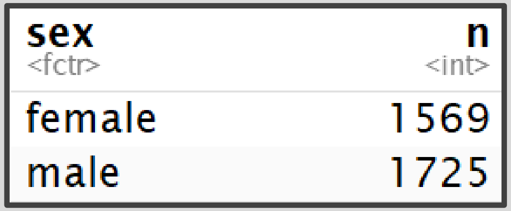
- Variation
  - Variable = Quantity, Quality, or Property You Can Measure
  - Reason: Values Tend to “Vary”
  - Example: Random
    - Categorical:
      - Eye Color
      - Occupation
    - Numerical:
      - Salary
      - Hair Count

Question

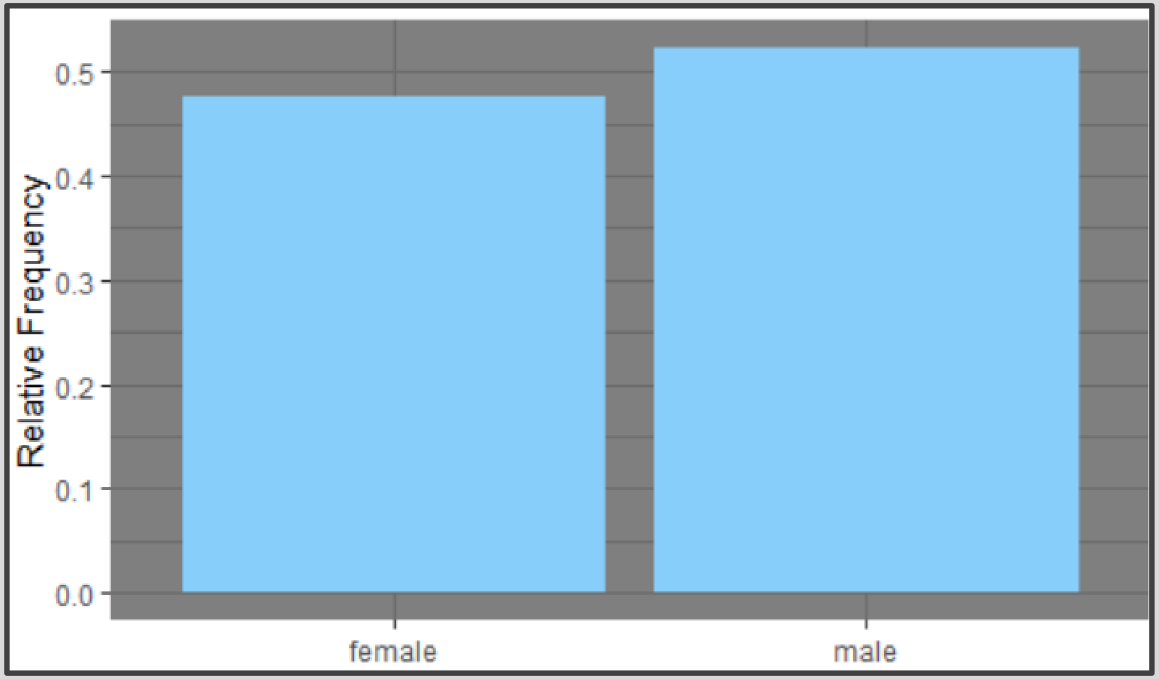
- Initial Questions
  - Example: Random
    - Which Eye Color Occurs Most Often?
    - Are Salaries Skewed?
    - Where is the Middle 50% of the Sample in Regards to Hair Count?
  - Example: Wages
    - What did the Workforce Look Like in Terms of Sex?
    - How Spread Out Were Wages in 1987?

Visualize
Summarize

- Variation Visualized
  - Example: Wages
    - Categorical: Sex

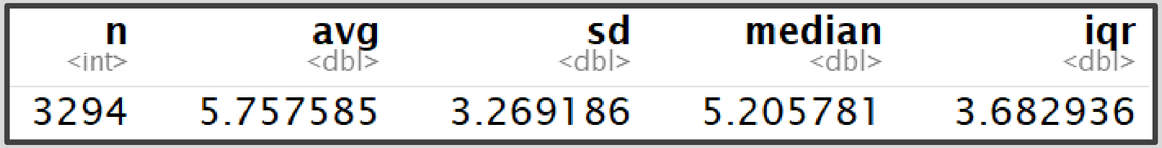


sex	n
<fctr>	<int>
female	1569
male	1725

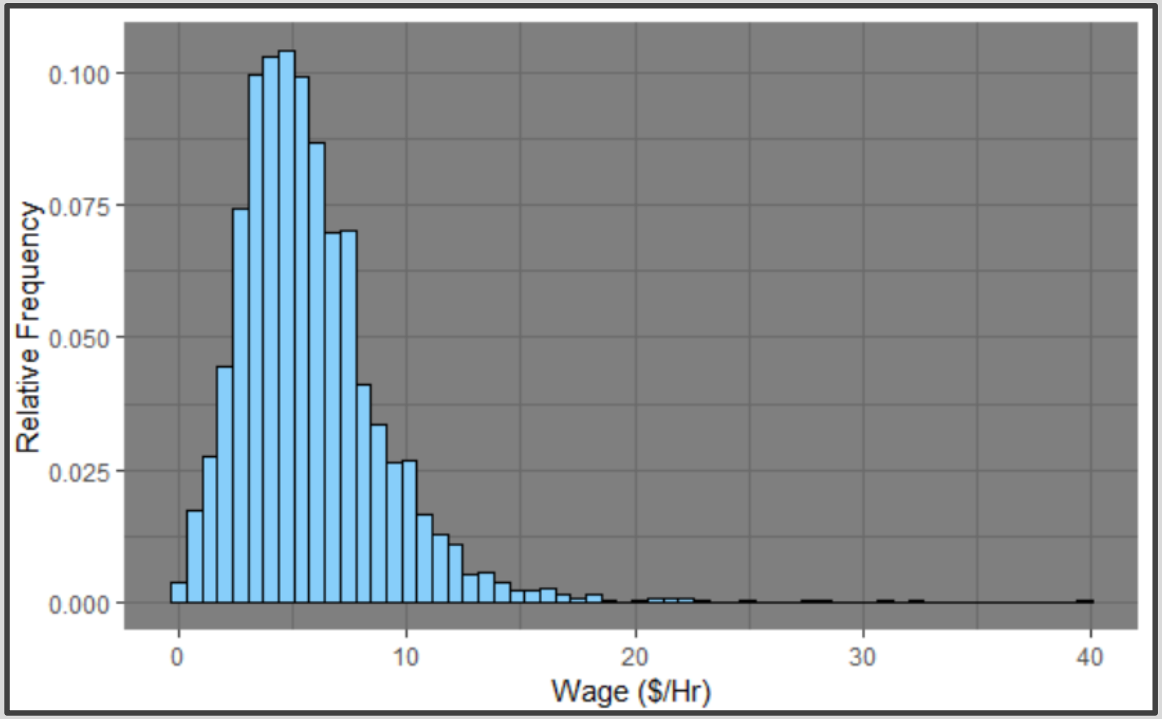


Visualize
Summarize

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



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

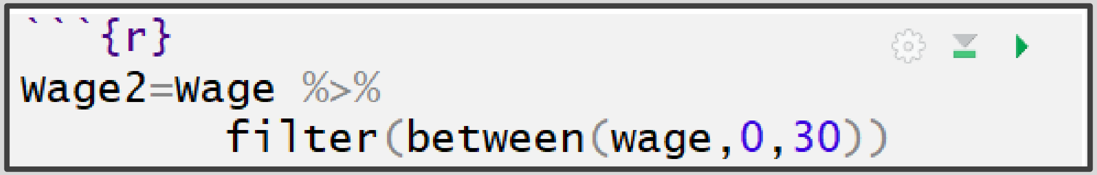


Unusual Values

- 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
- Problems:
  - Book: Visualization
  - Other: Inference

Unusual Values

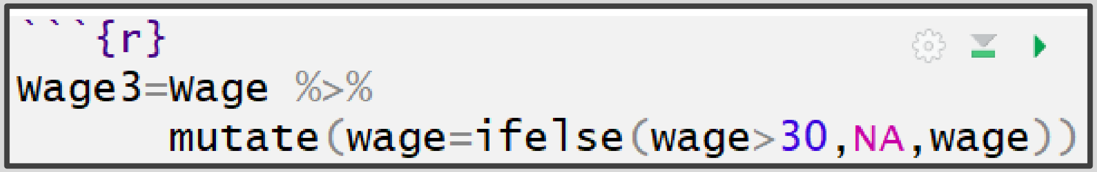
- Example: Wages
  - Few People Above 30 $/Hr
  - Drop Entire Row



~~~
```{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

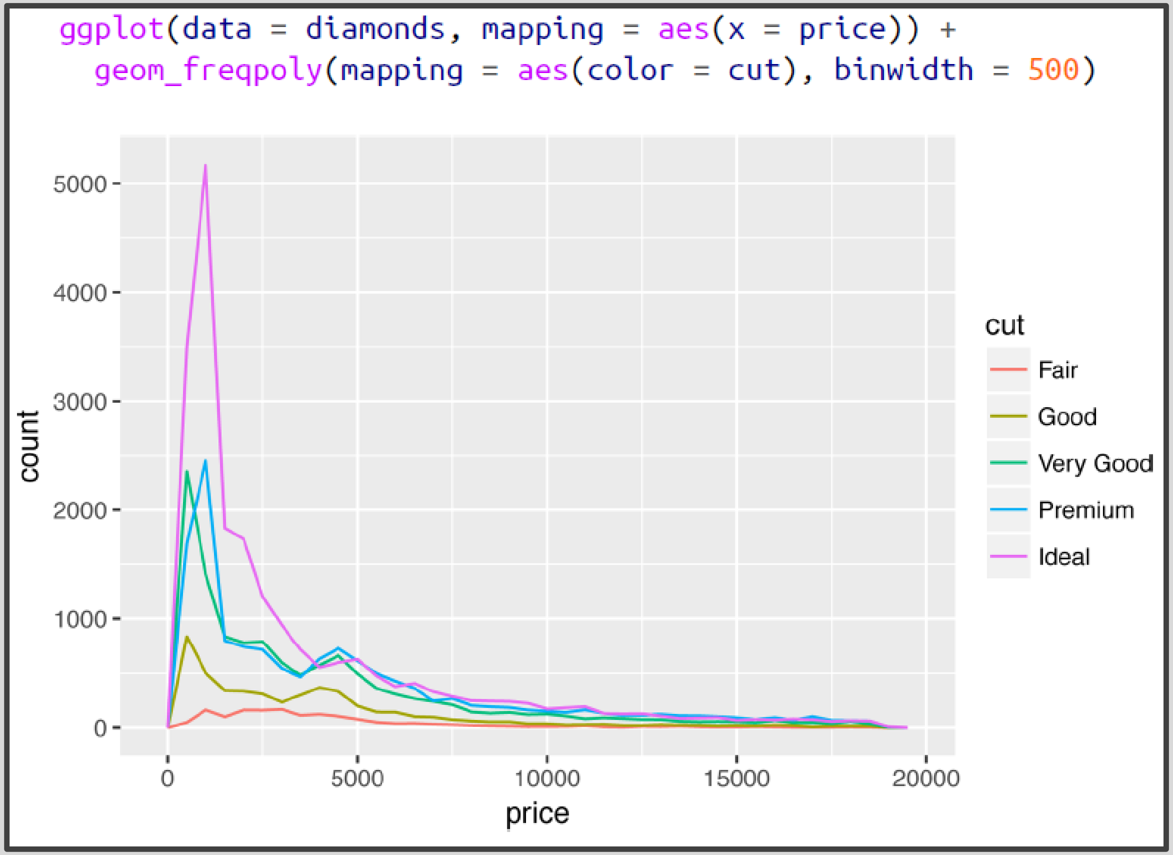
- Covariation
  - Goal: Explain Variation
  - Describes the Behavior Between Variables
  - We Often Attempt to Explain Variation **Within** by Looking at Covariation **Between**
  - Identify the **Signal** despite the **Noise**

Question

- More Questions
  - Example: Random
    - Are there Occupations with an Unusual Distribution of Eye Color?
    - Does Occupation Affect Salary?
    - What is the Relationship Between Salary and Hair Count?
  - Example: Wages

Visualize
Summarize

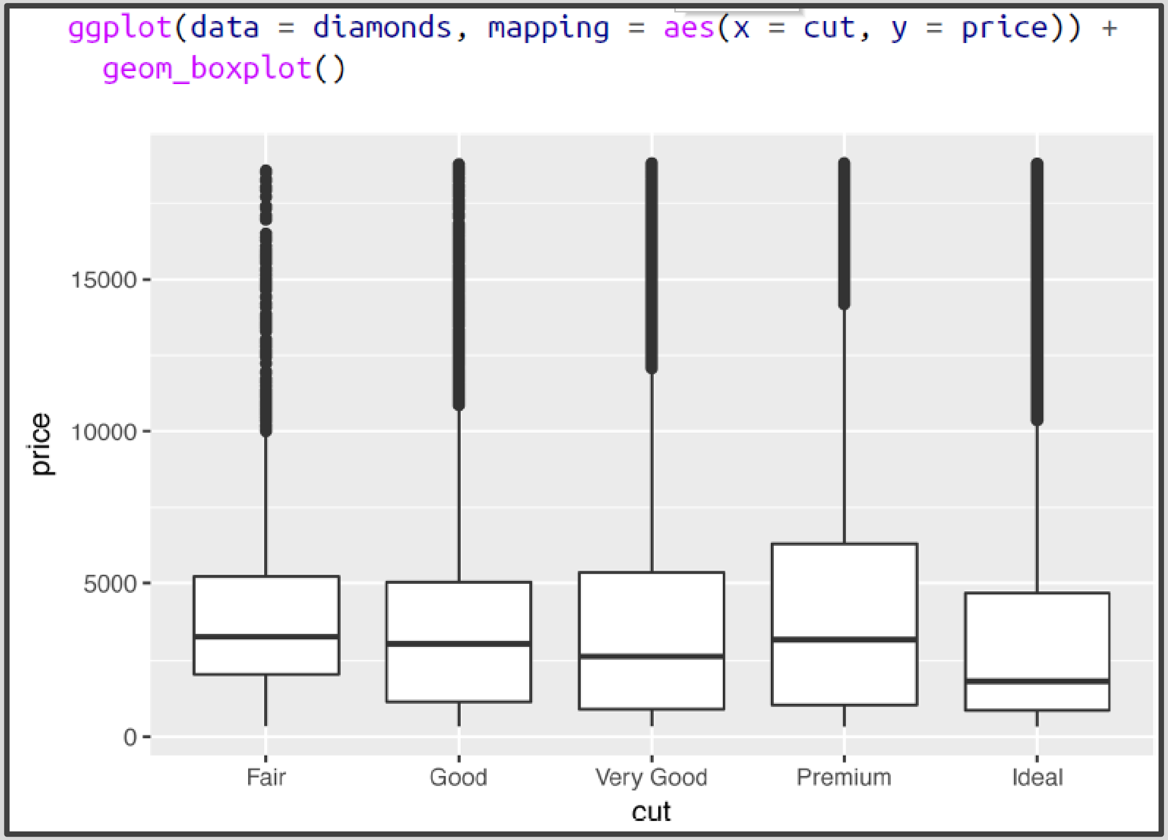
- Categorical and Numeric



```
ggplot(data = diamonds, mapping = aes(x = price)) +
  geom_freqpoly(mapping = aes(color = cut), binwidth = 500)
```

Visualize
Summarize

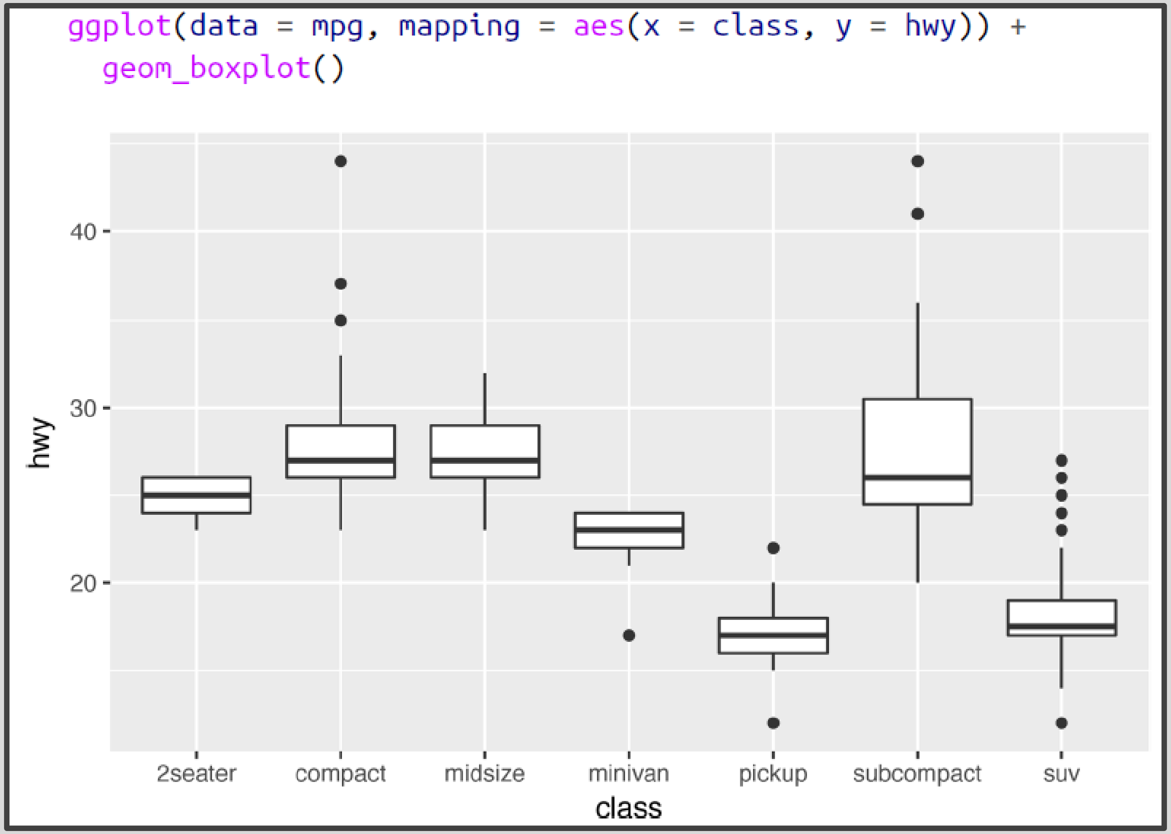
- Categorical and Numeric



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

Visualize Summarize

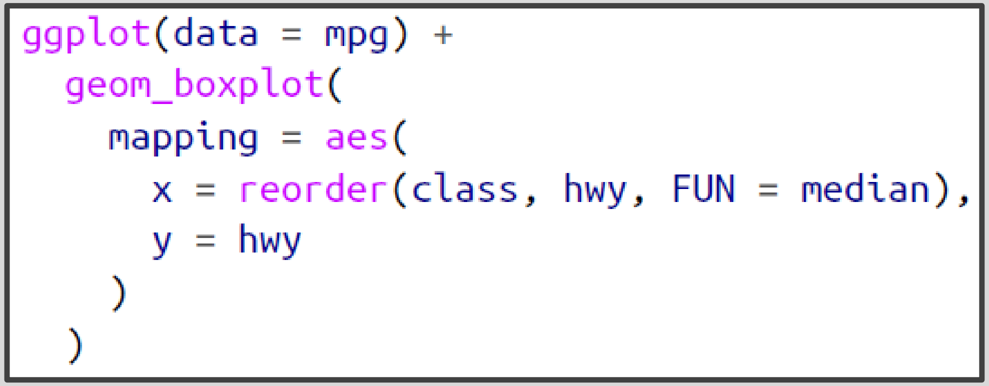
- Categorical and Numeric



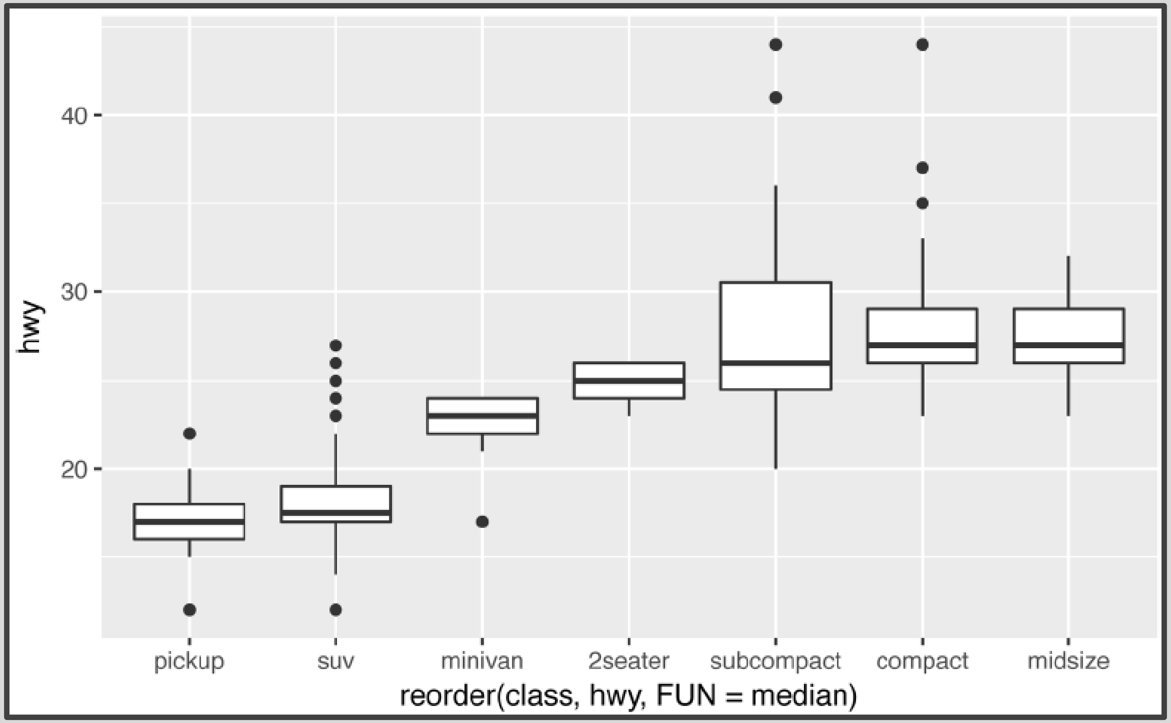
```
ggplot(data = mpg, mapping = aes(x = class, y = hwy)) +
  geom_boxplot()
```

Visualize
Summarize

- Categorical and Numeric

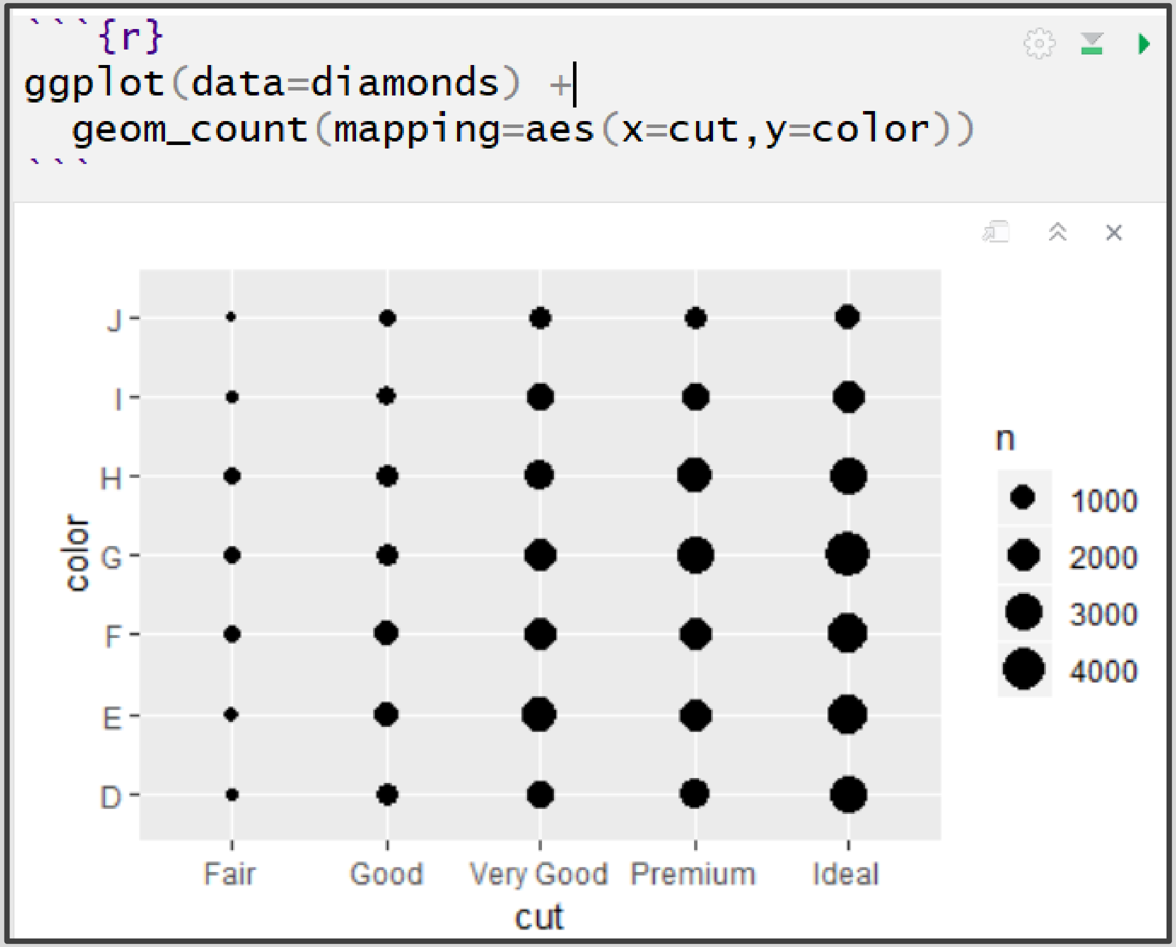


```
ggplot(data = mpg) +
  geom_boxplot(
    mapping = aes(
      x = reorder(class, hwy, FUN = median),
      y = hwy
    )
  )
```



Visualize
Summarize

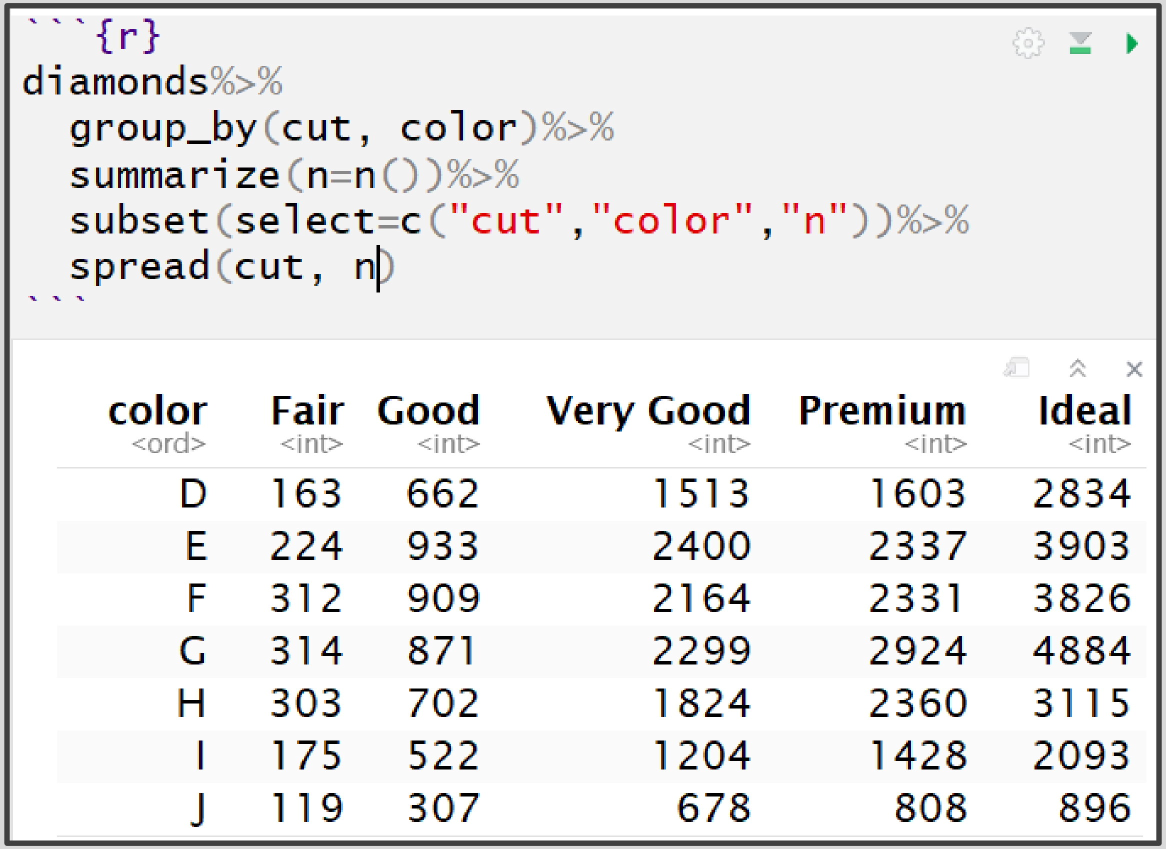
- Categorical and Categorical



~~~
```{r}
ggplot(data=diamonds) +
  geom_count(mapping=aes(x=cut,y=color))
```
~~~

Visualize
Summarize

- Categorical and Categorical

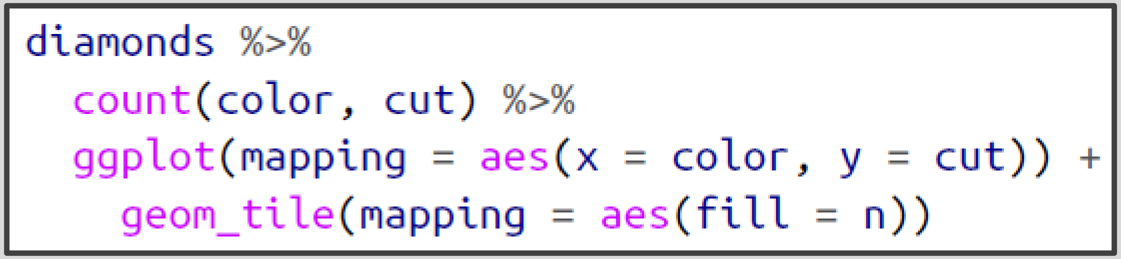


~~~
```{r}
diamonds%>%
  group_by(cut, color)%>%
  summarize(n=n())%>%
  subset(select=c("cut","color","n"))%>%
  spread(cut, n)
```
~~~

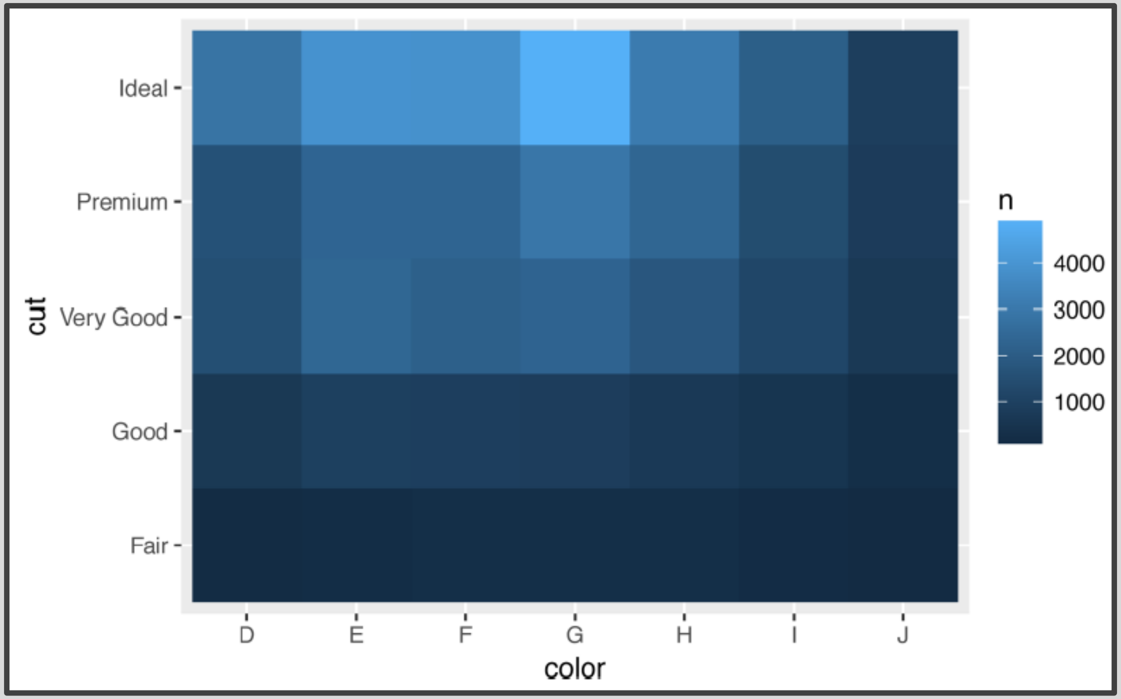
color <ord>	Fair <int>	Good <int>	Very Good <int>	Premium <int>	Ideal <int>
D	163	662	1513	1603	2834
E	224	933	2400	2337	3903
F	312	909	2164	2331	3826
G	314	871	2299	2924	4884
H	303	702	1824	2360	3115
I	175	522	1204	1428	2093
J	119	307	678	808	896

Visualize
Summarize

- Categorical and Categorical

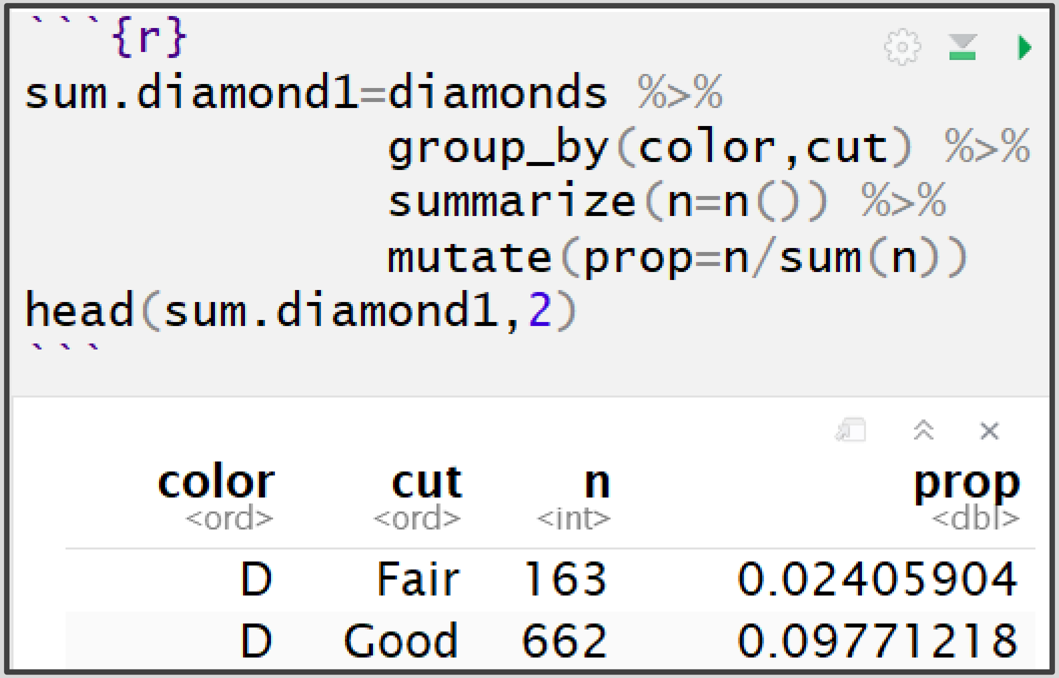


```
diamonds %>%
  count(color, cut) %>%
  ggplot(mapping = aes(x = color, y = cut)) +
    geom_tile(mapping = aes(fill = n))
```



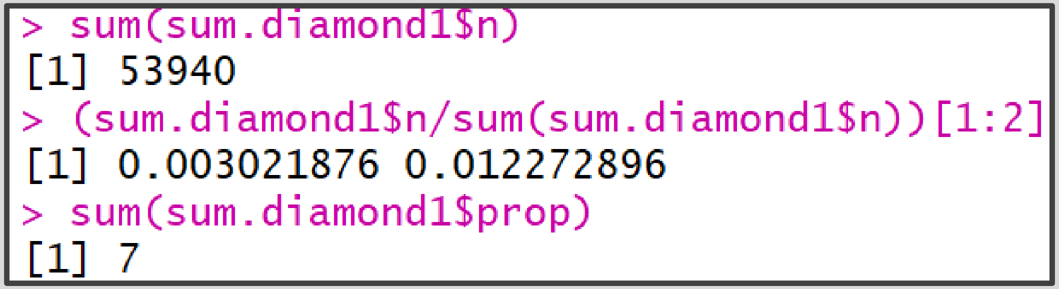
Visualize Summarize

- Categorical and Categorical



~~~
```{r}
sum.diamond1=diamonds %>%
                group_by(color,cut) %>%
                summarize(n=n()) %>%
                mutate(prop=n/sum(n))
head(sum.diamond1,2)
```
~~~

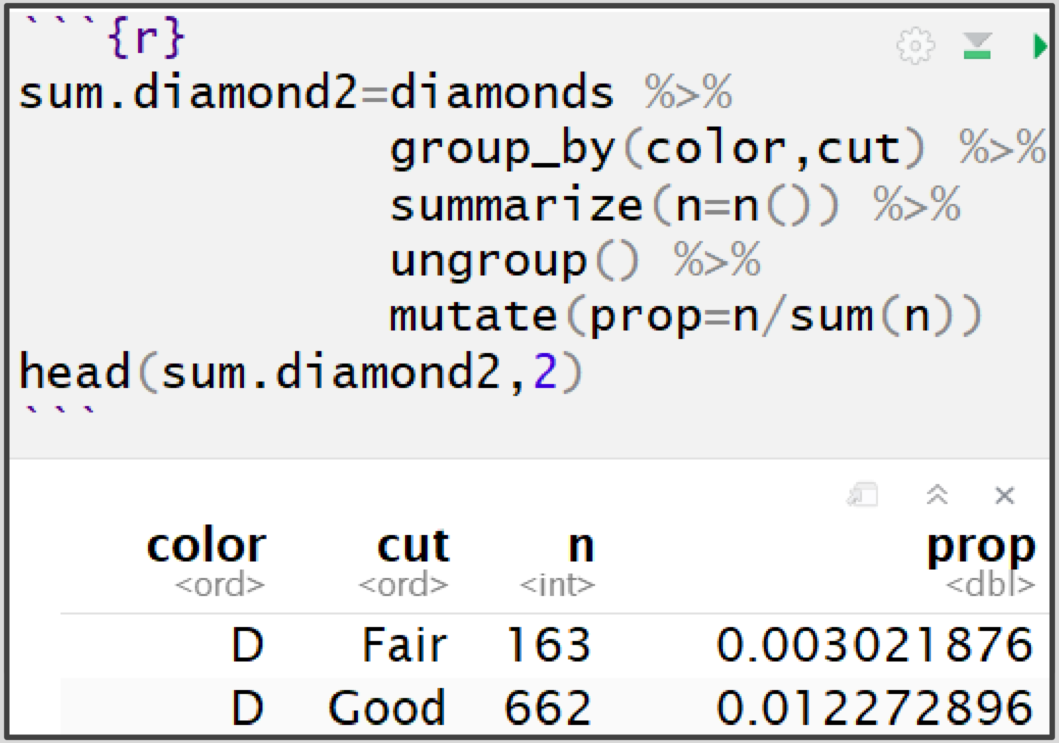
color <ord>	cut <ord>	n <int>	prop <dbl>
D	Fair	163	0.02405904
D	Good	662	0.09771218



```
> sum(sum.diamond1$n)
[1] 53940
> (sum.diamond1$n/sum(sum.diamond1$n))[1:2]
[1] 0.003021876 0.012272896
> sum(sum.diamond1$prop)
[1] 7
```

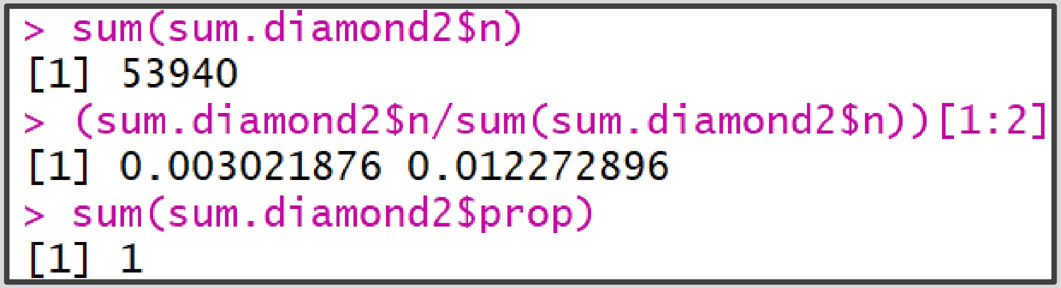
Visualize
Summarize

- Categorical and Categorical



~~~
```{r}
sum.diamond2=diamonds %>%
                group_by(color,cut) %>%
                summarize(n=n()) %>%
                ungroup() %>%
                mutate(prop=n/sum(n))
head(sum.diamond2,2)
```
~~~

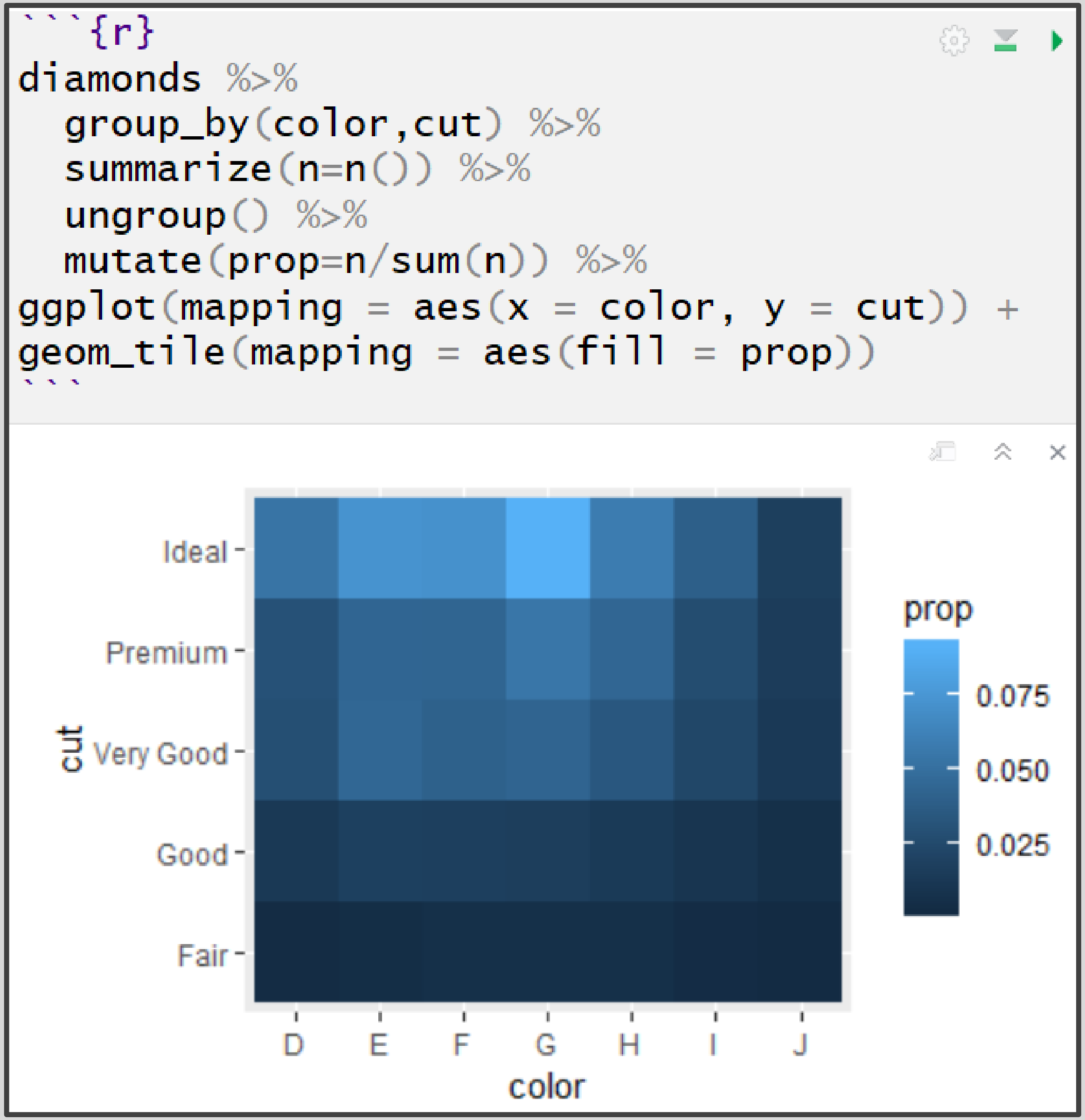
color <ord>	cut <ord>	n <int>	prop <dbl>
D	Fair	163	0.003021876
D	Good	662	0.012272896



```
> sum(sum.diamond2$n)
[1] 53940
> (sum.diamond2$n/sum(sum.diamond2$n))[1:2]
[1] 0.003021876 0.012272896
> sum(sum.diamond2$prop)
[1] 1
```

Visualize Summarize

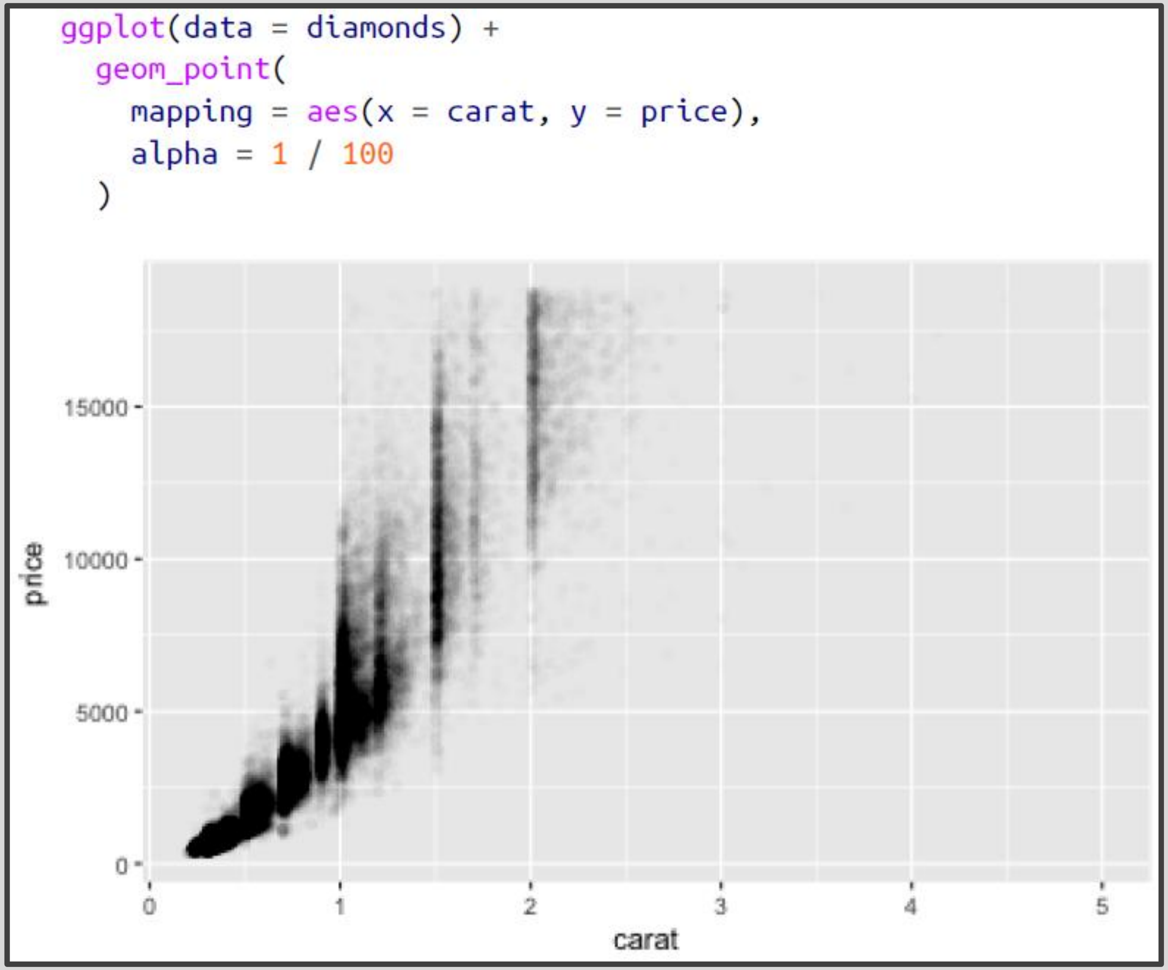
- Categorical and Categorical



~~~
```{r}
diamonds %>%
  group_by(color,cut) %>%
  summarize(n=n()) %>%
  ungroup() %>%
  mutate(prop=n/sum(n)) %>%
ggplot(mapping = aes(x = color, y = cut)) +
geom_tile(mapping = aes(fill = prop))
```
~~~

Visualize
Summarize

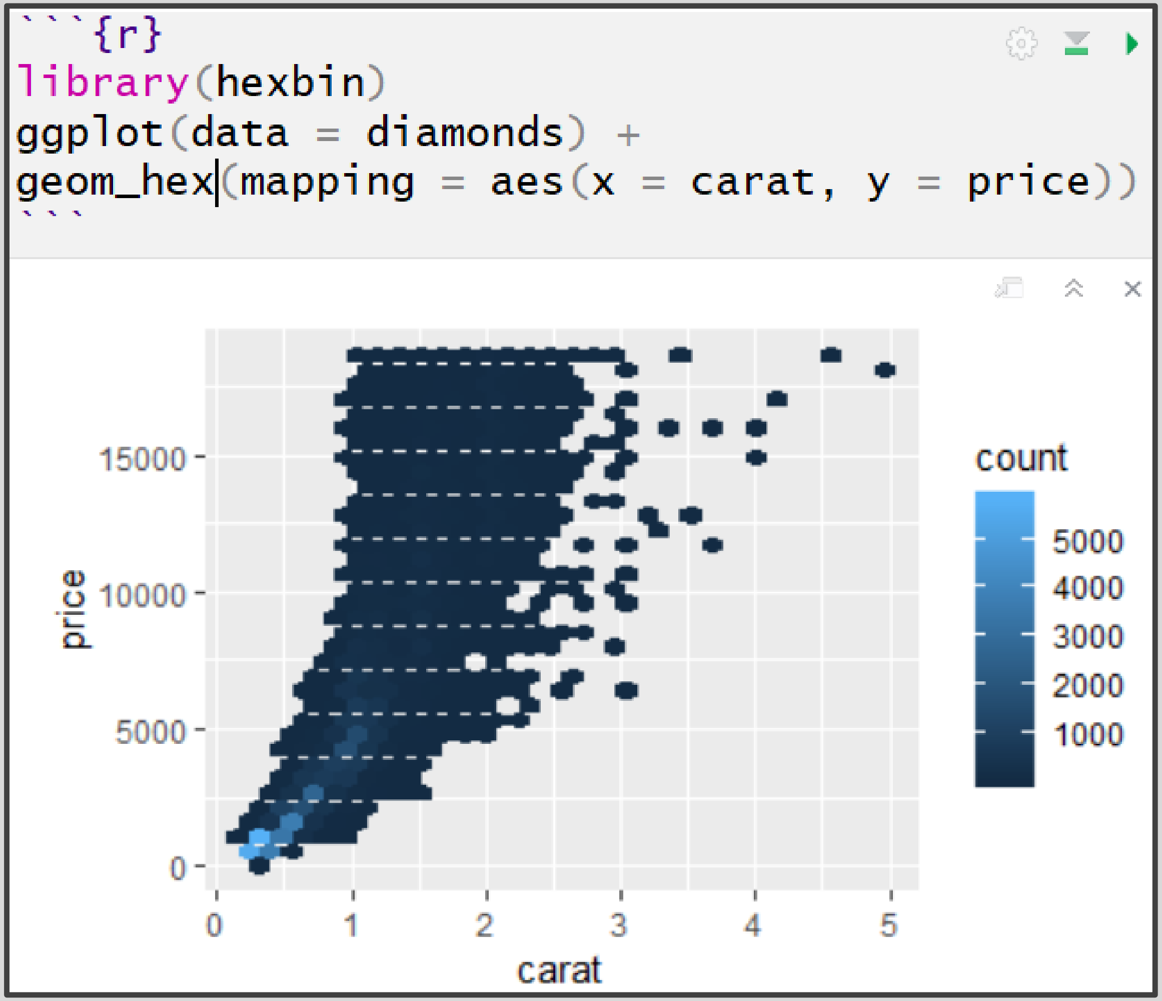
- Numerical and Numerical



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

Visualize
Summarize

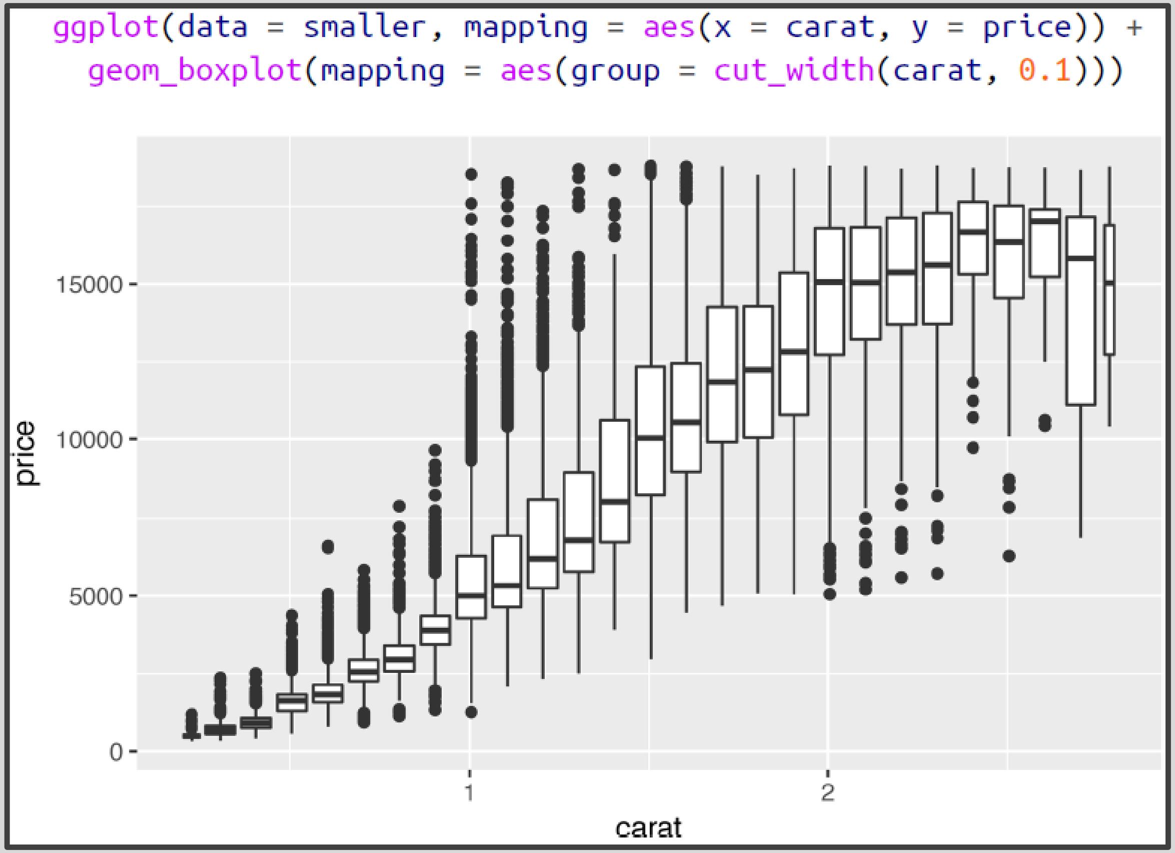
- Numerical and Numerical



~~~
```{r}
library(hexbin)
ggplot(data = diamonds) +
geom_hex(mapping = aes(x = carat, y = price))
```
~~~

Visualize
Summarize

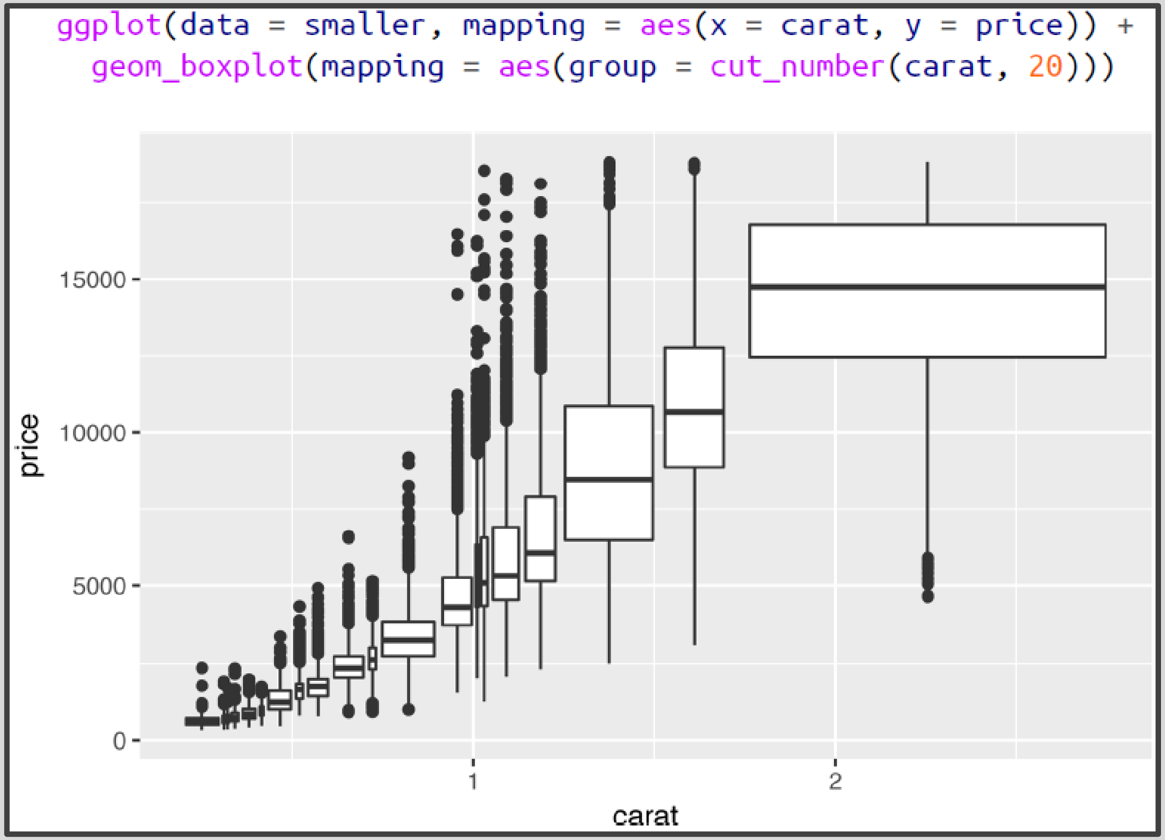
- Numerical and Numerical



```
ggplot(data = smaller, mapping = aes(x = carat, y = price)) +
  geom_boxplot(mapping = aes(group = cut_width(carat, 0.1)))
```

Visualize
Summarize

- Numerical and Numerical



```
ggplot(data = smaller, mapping = aes(x = carat, y = price)) +
  geom_boxplot(mapping = aes(group = cut_number(carat, 20)))
```

Closing

Disperse and Make Reasonable Decisions