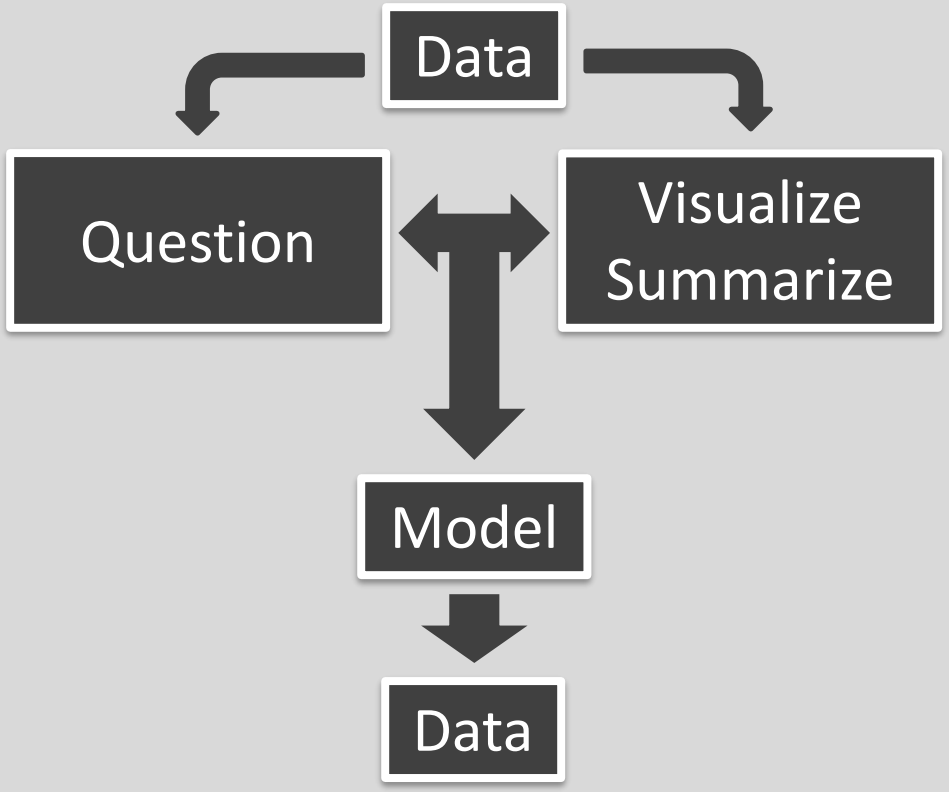
Exploratory Data Analysis II

EDA Defined

- Remember the Process



- Be the Process

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?

EDA Mantra

"Be the change that you want to see in the world unless that change is statistically insignificant."

- Mahatma Mario

Question

What is the relationship between

the size of the

and

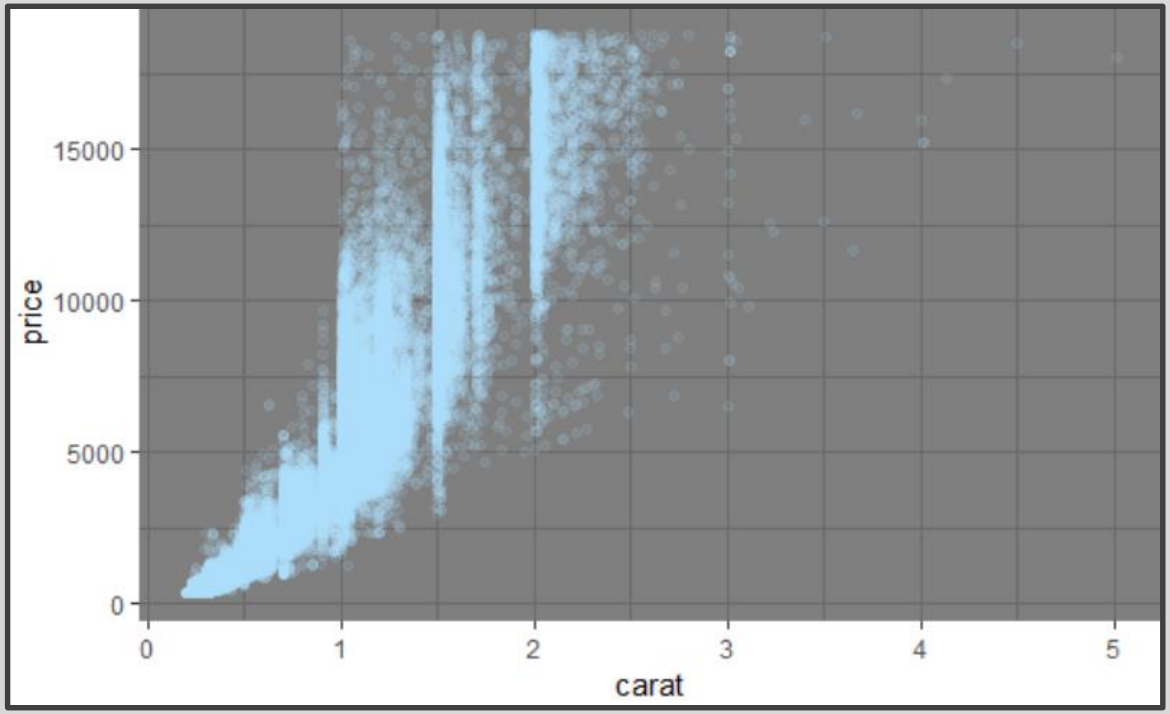
the price of the

?

Visualize
Summarize

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

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

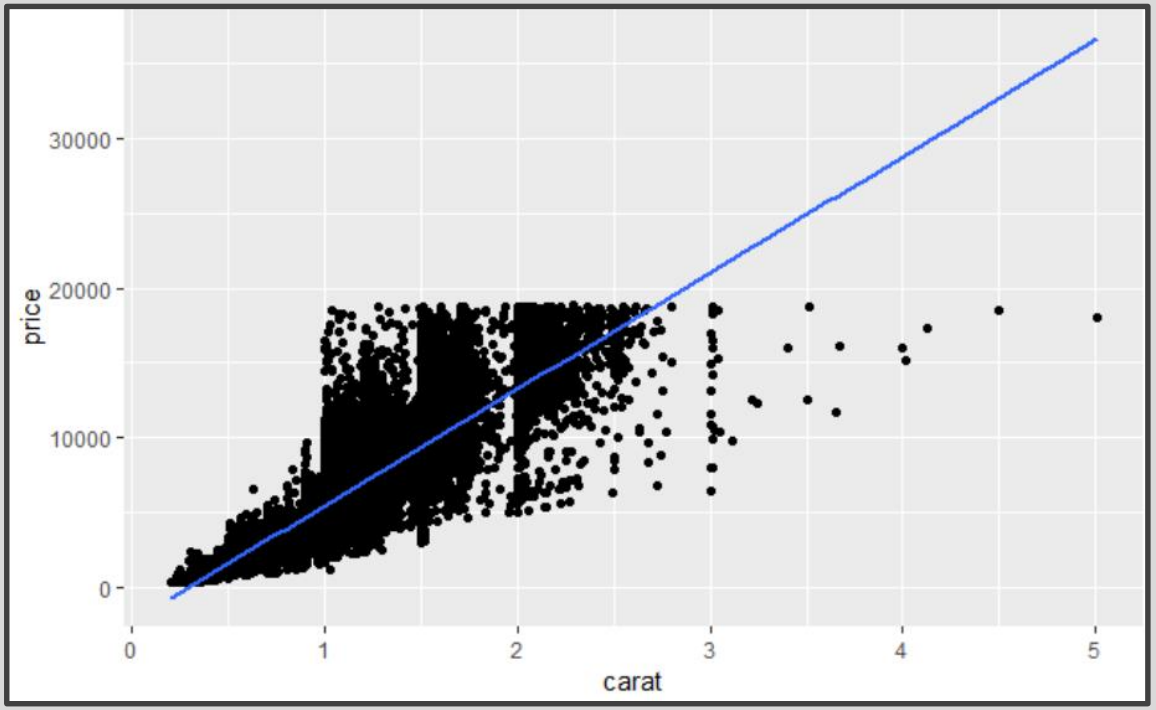


Question

- 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

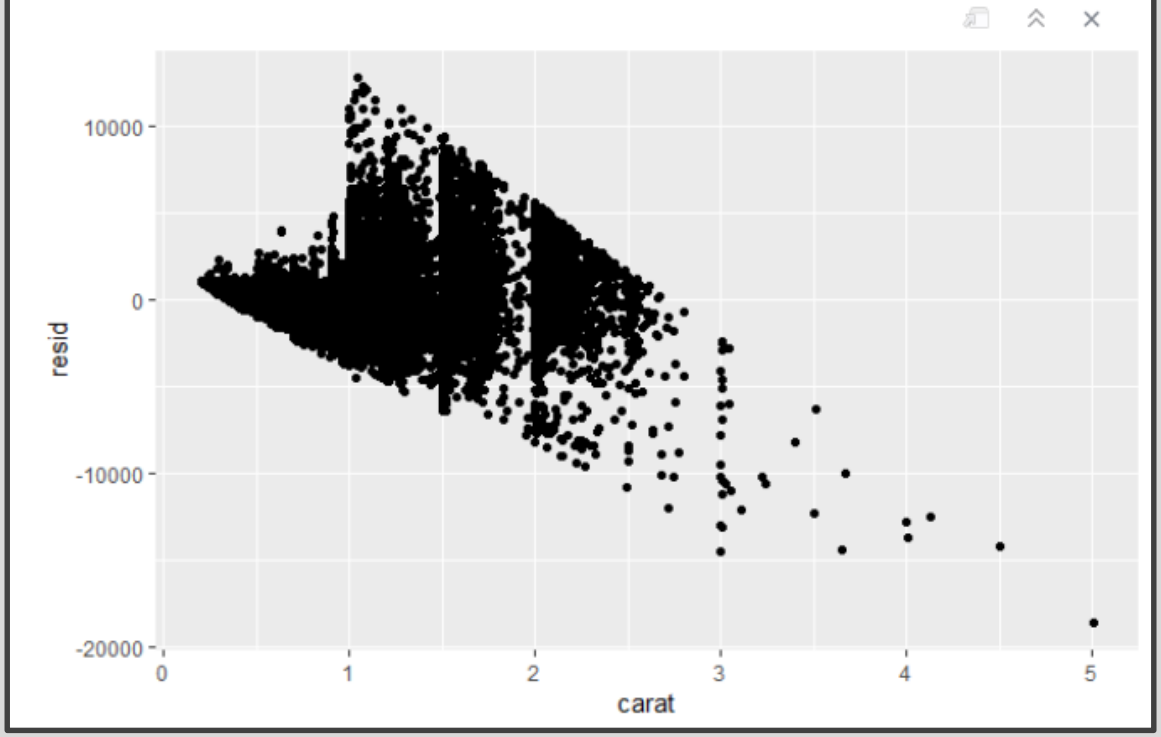
- Linear Model



Model

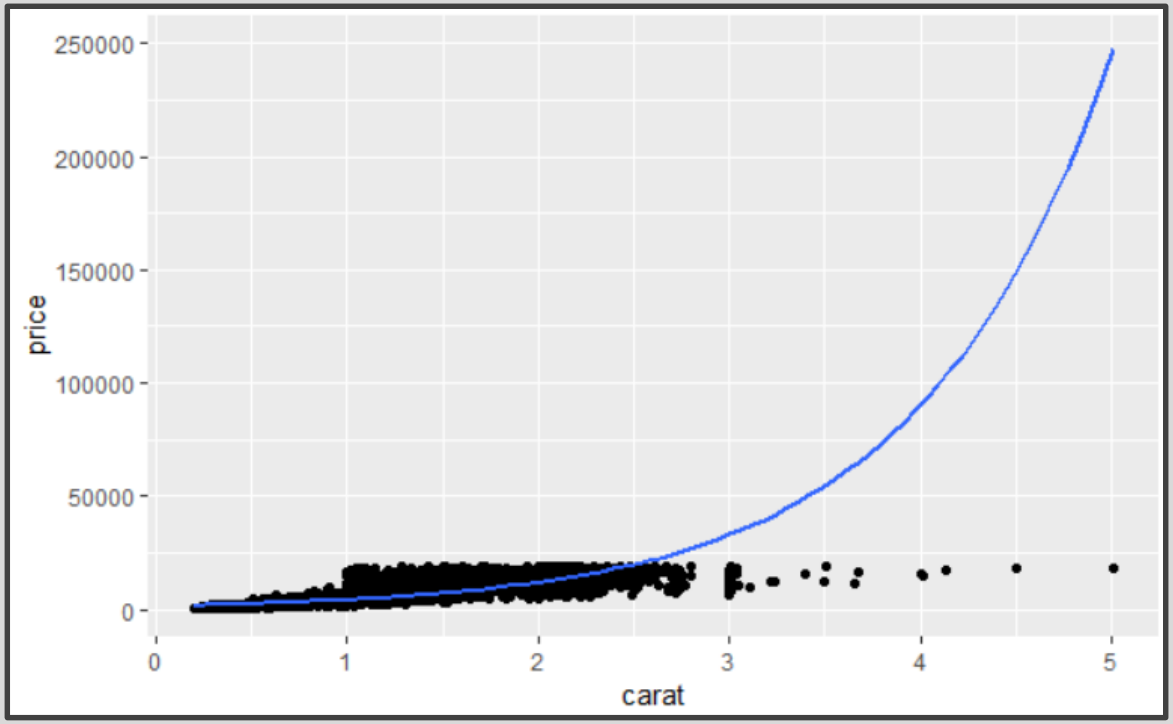
- Linear Model

```r
library(modelr)
lin.mod=lm(price~carat,data=diamonds)
diamonds.lin.resid = diamonds %>%
  add_residuals(mod=lin.mod)
ggplot(data=diamonds.lin.resid) +
  geom_point(aes(x=carat,y=resid))
```



Model

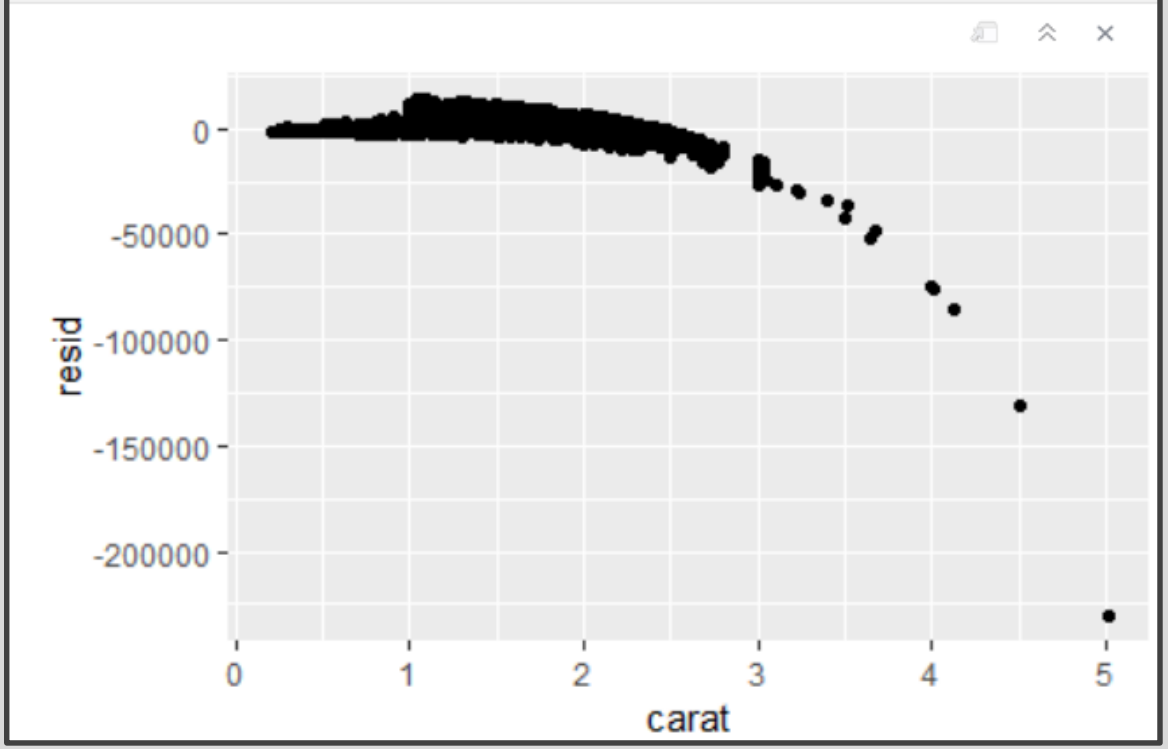
- Exponential Model



Model

- Exponential Model

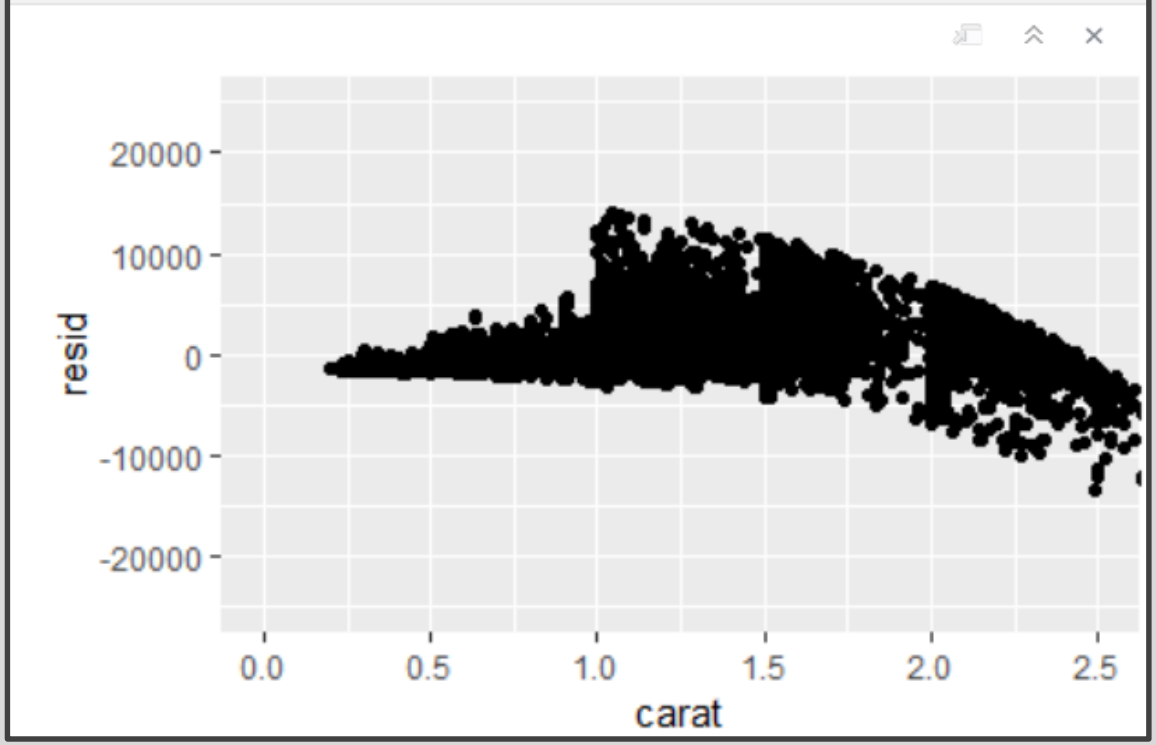
```r
exp.mod=lm(price~exp(carat),data=diamonds)
diamonds.exp.resid = diamonds %>%
  add_residuals(mod=exp.mod)
ggplot(data=diamonds.exp.resid) +
  geom_point(aes(x=carat,y=resid))
```



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))
```



Closing

Disperse and Make Reasonable Decisions