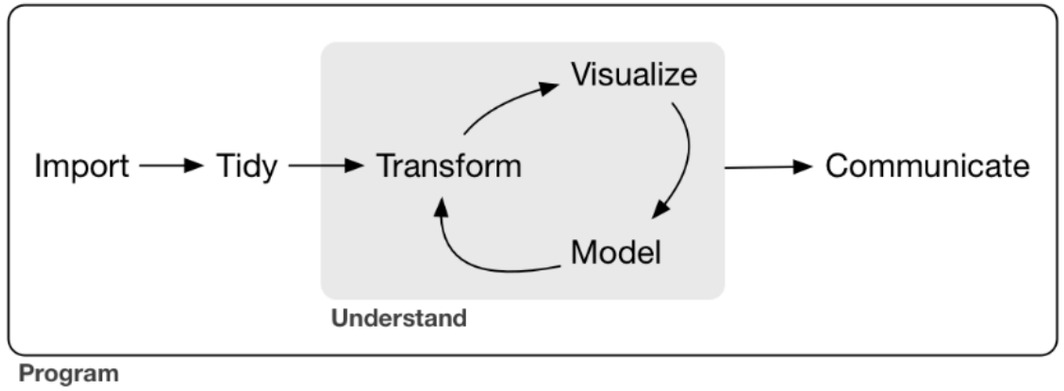


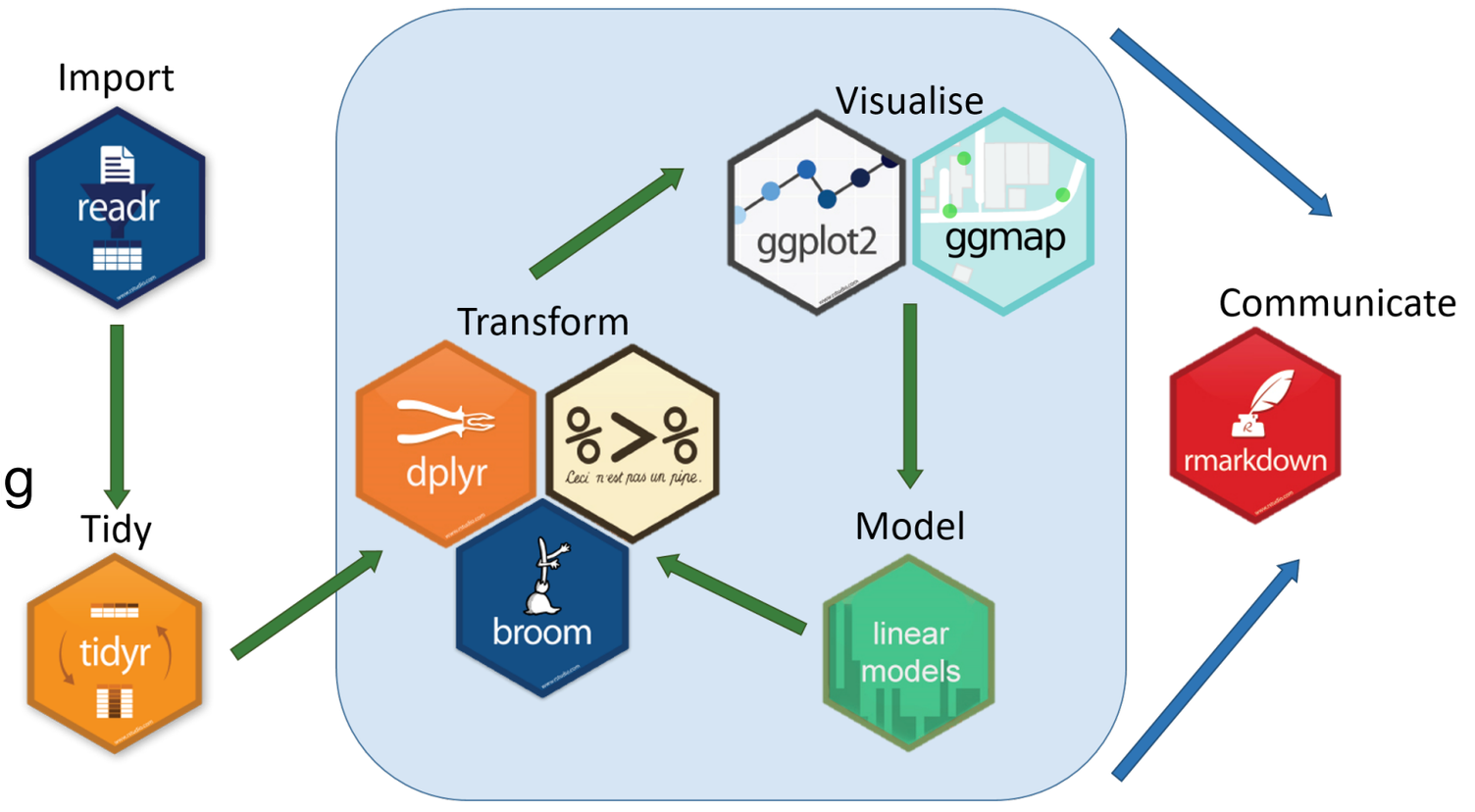
Data Science: Tidyverse

Alex Di Genova

14/05/2024

What is Tidyverse?

- A collection of R packages for data science
  - ggplot2 — data visualization
  - dplyr — data manipulation
  - tidyr — data tidying
  - readr — data import
  - purrr — functional programming
  - tibble — modern dataframes
  - stringr — string manipulation
  - forcats — factor handling

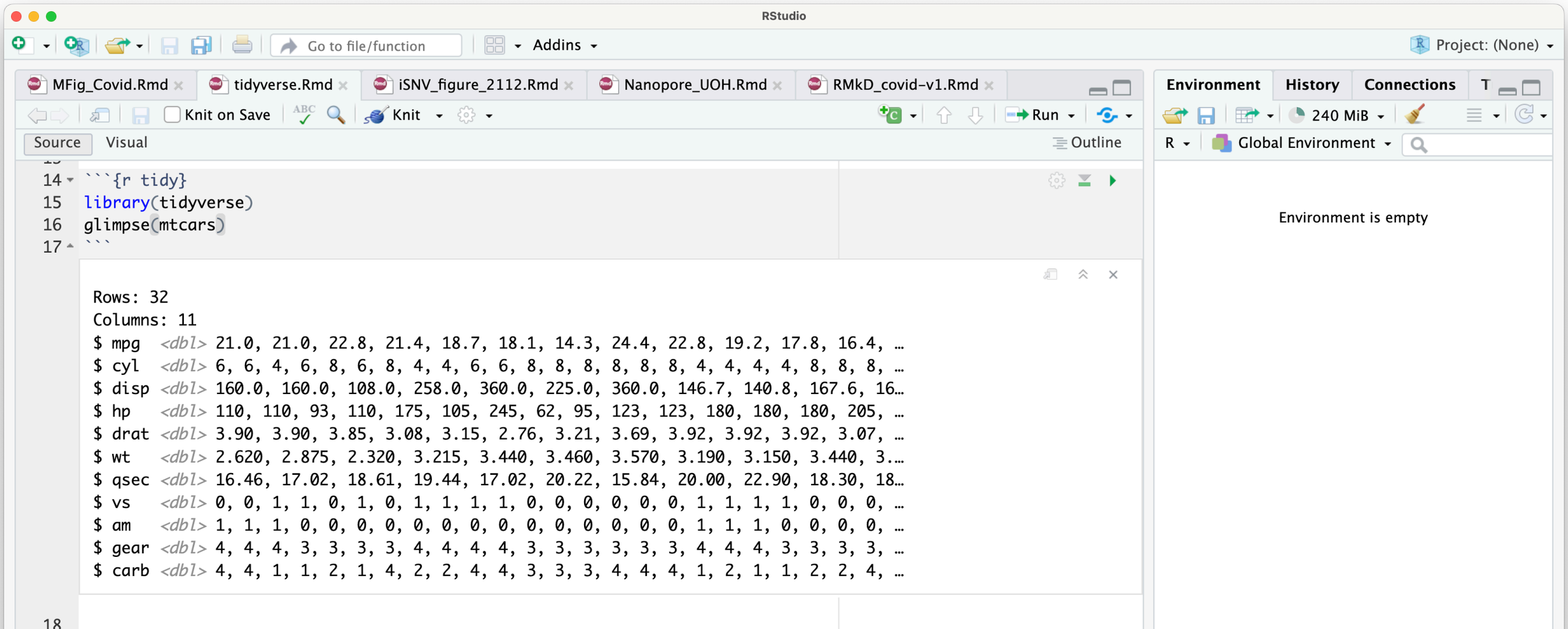


- Data science workflow (import, clean, transform, visualize, model)

Tidyverse

Data exploration

- install.packages(“tidyverse")
- glimpse(mtcars)
- summary(mtcars)



Tidyverse

In action

- Input is always a dataframe
- Each row is an observation and each column a single variable
- The pipe %>% operator, guides the flow operations of data.
- Linux + SQL

```
18
19 ## tidyverse
20
21 operators and verbs
22
23 ```{r pressure, echo=FALSE}
24 iris %>%
25   select(Petal.Length, Petal.Width, Species) %>%
26   filter(Species %in% c("versicolor", "setosa")) %>%
27   group_by(Species) %>%
28   summarize(
29     AvgPetalLength = mean(Petal.Length),
30     AvgPetalWidth = mean(Petal.Width))
31 ```
```

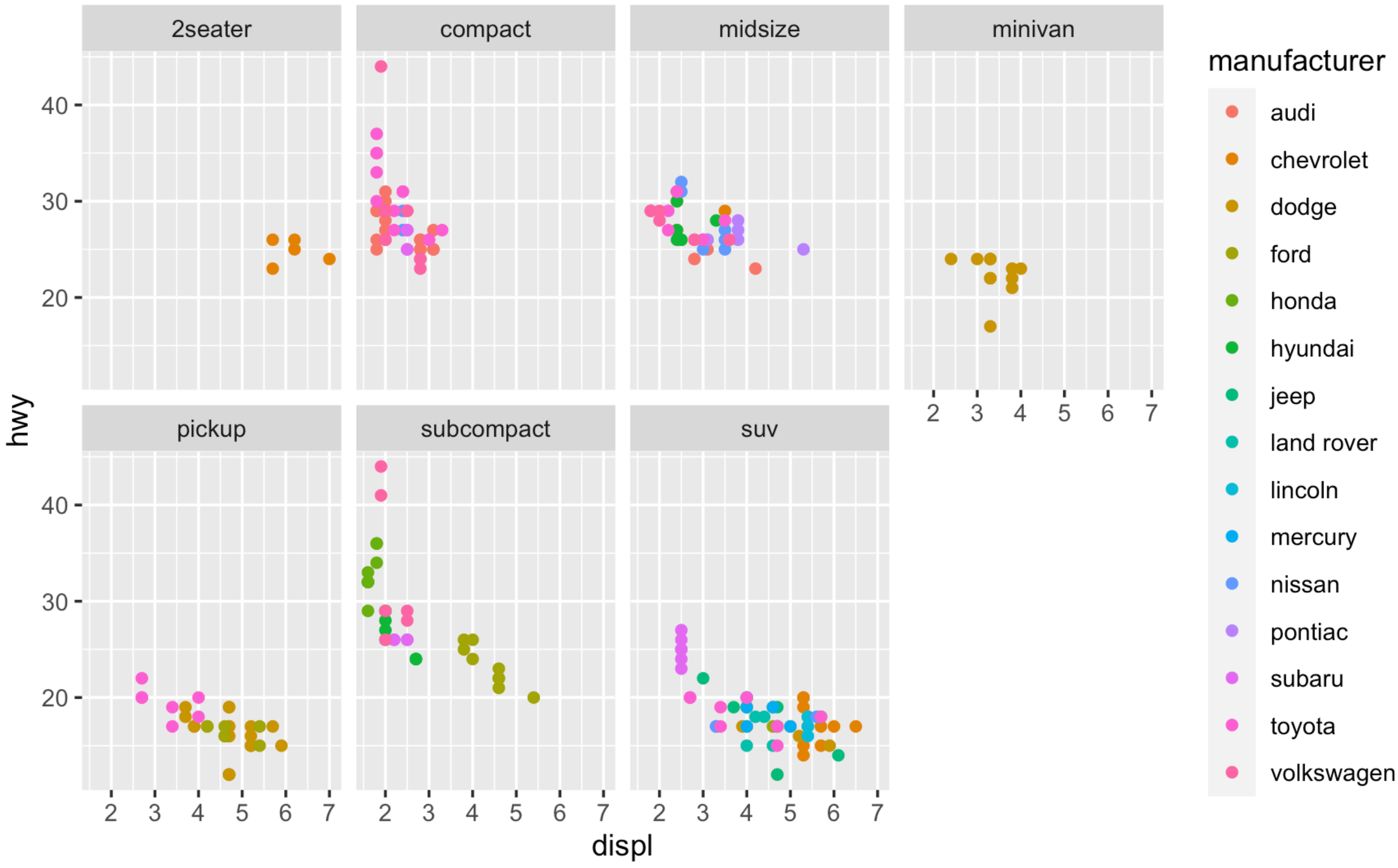
A tibble: 2 × 3

Species <fctr>	AvgPetalLength <dbl>	AvgPetalWidth <dbl>
setosa	1.462	0.246
versicolor	4.260	1.326

Facets

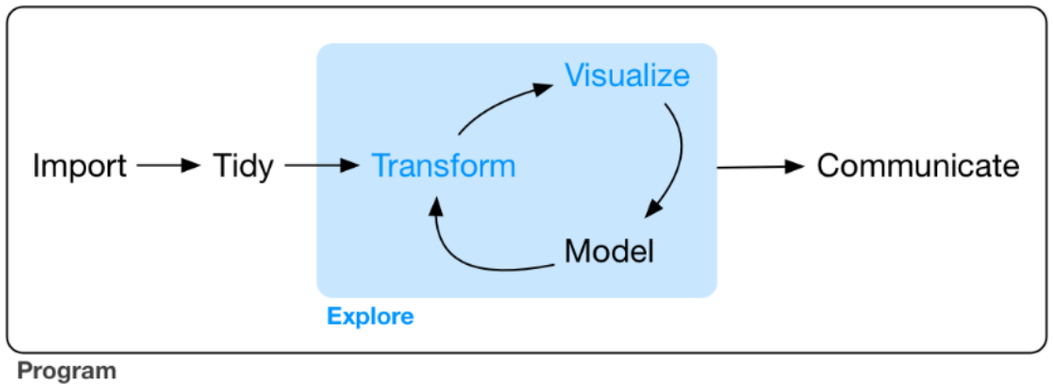
```{r ggplot2}
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy, color=manufacturer)) +
  facet_wrap(~ class, nrow = 2)

```



Tidyverse

Explore



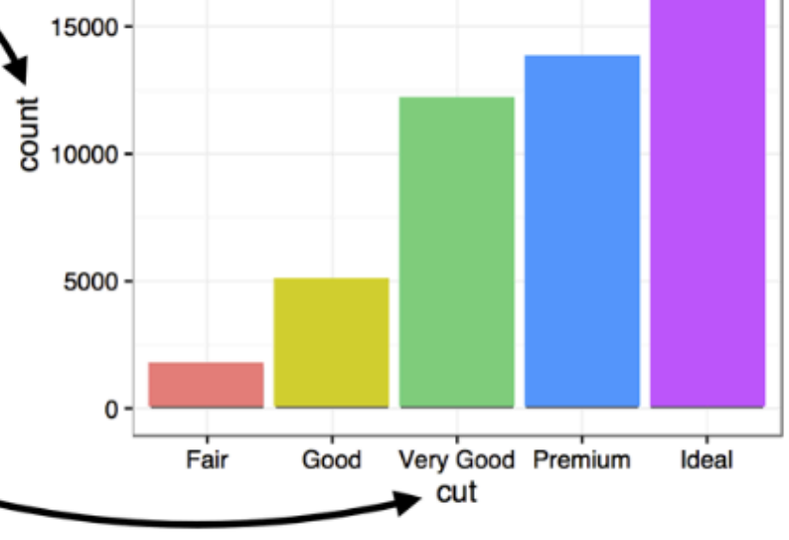
carat	cut	color	clarity	depth	table	price	x	y	z
0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
0.29	Premium	I	VS2	62.4	58	334	4.20	4.23	2.63
0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
...	...	...	...	...	...	...	...	...	...

stat_count()

cut	count	prop
Fair	1610	1
Good	4906	1
Very Good	12082	1
Premium	13791	1
Ideal	21551	1

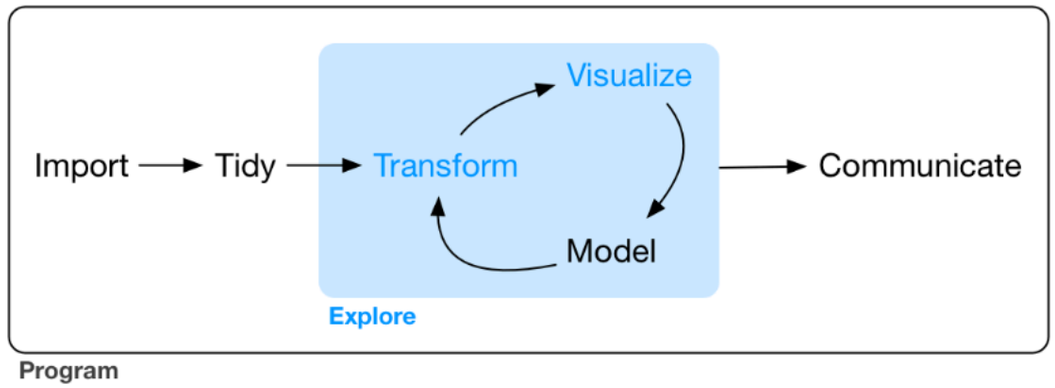
5. Place geoms in a cartesian coordinate system.

6. Map the y values to **..count..** and the x values to **cut**.



Tidyverse

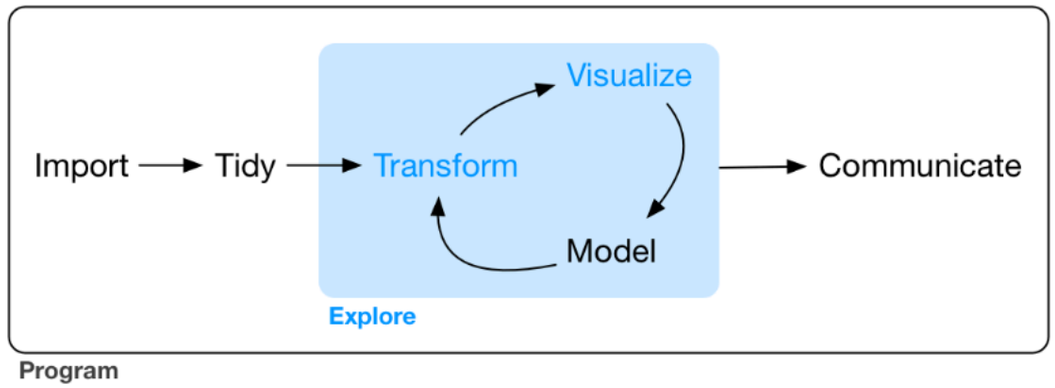
Dpylr — transform



- Data manipulation challenges:
  - Pick observations by their values (filter())
  - Reorder the rows (arrange())
  - Pick variables by their names (select())
  - Create new variables (mutate())
  - Collapse many values to a single summary (summarize())
  - Group rows (group_by())
- All dplyr verbs expect a data.frame and produce a new data.frame

Tidyverse

Dpylr — transform



- filter()
  - R provides the standard suite: >, >=, <, <=, != (not equal), and == (equal)
  - Log | y

```
## Dplyr

### Filter

```{r filter}
mpg %>% filter(manufacturer == "audi") %>% head()
```
```

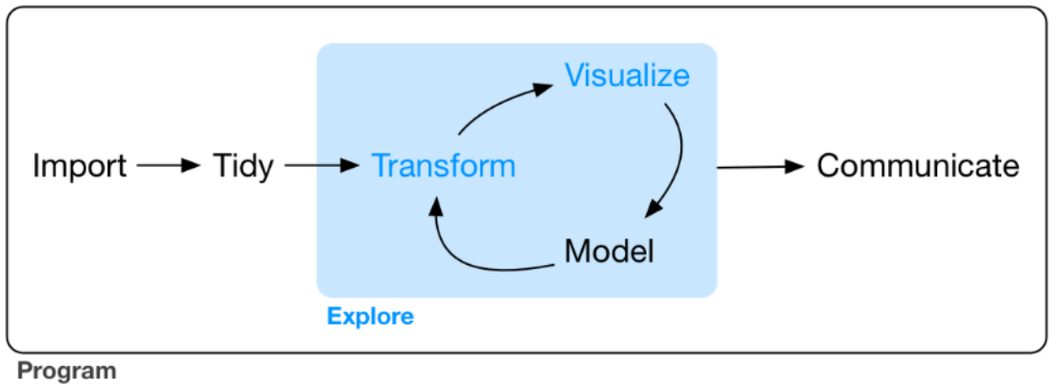
A tibble: 6 × 11

manufacturer <chr>	model <chr>	displ <dbl>	year <int>	cyl <int>	trans <chr>	drv <chr>	cty <int>	hwy <int>	fl <chr>
audi	a4	1.8	1999	4	auto(l5)	f	18	29	p
audi	a4	1.8	1999	4	manual(m5)	f	21	29	p
audi	a4	2.0	2008	4	manual(m6)	f	20	31	p
audi	a4	2.0	2008	4	auto(av)	f	21	30	p
audi	a4	2.8	1999	6	auto(l5)	f	16	26	p
audi	a4	2.8	1999	6	manual(m5)	f	18	26	p

6 rows | 1-10 of 11 columns

Tidyverse

Dpylr — transform




```{r filter2}
mpg %>% filter(manufacturer %in% c("audi","chevrolet")) %>% arrange(desc(year),manufacturer)
```

A tibble: 37 × 11

manufacturer <chr>	model <chr>	displ <dbl>	year <int>	cyl <int>	trans <chr>	drv <chr>	cty <int>	h... <int>	fl <chr>
audi	a4	2.0	2008	4	manual(m6)	f	20	31	p
audi	a4	2.0	2008	4	auto(av)	f	21	30	p
audi	a4	3.1	2008	6	auto(av)	f	18	27	p
audi	a4 quattro	2.0	2008	4	manual(m6)	4	20	28	p
audi	a4 quattro	2.0	2008	4	auto(s6)	4	19	27	p
audi	a4 quattro	3.1	2008	6	auto(s6)	4	17	25	p
audi	a4 quattro	3.1	2008	6	manual(m6)	4	15	25	p
audi	a6 quattro	3.1	2008	6	auto(s6)	4	17	25	p
audi	a6 quattro	4.2	2008	8	auto(s6)	4	16	23	p
chevrolet	c1500 suburban 2wd	5.3	2008	8	auto(l4)	r	14	20	r

1–10 of 37 rows | 1–10 of 11 columns

Previous 1 2 3 4 Next

Rmarkdown

Blog