EPSY 5261 : Introductory Statistical Methods

Day 4

Visualization and Numerical Summaries

Learning Goals

- At the end of this lesson, you should be able to…
  - Describe the general structure of a data set
    - What is a row?
    - What is a column?
  - Identify an attribute as categorical or quantitative
  - Create graphs for each variable type
  - Create summary statistics for each variable type

Data Structure

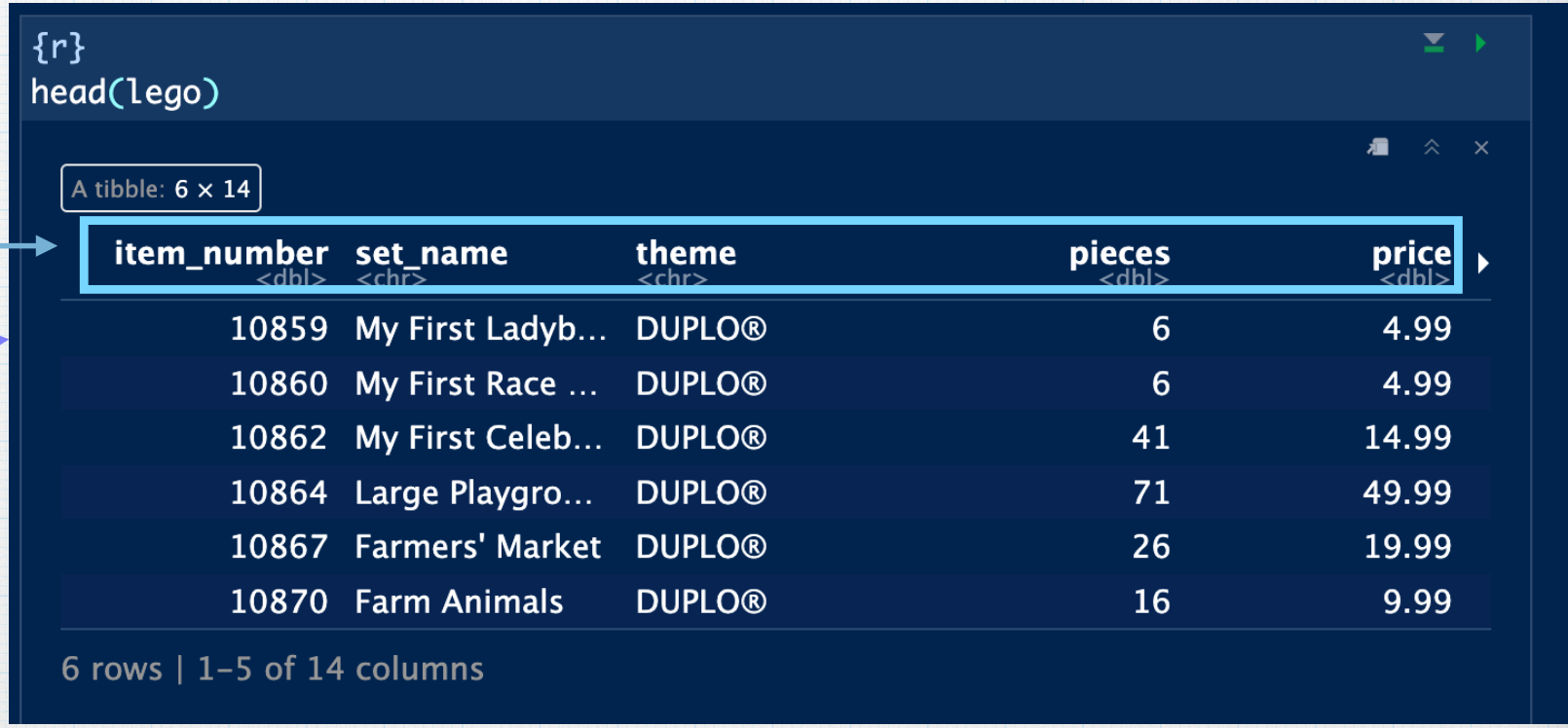
Data Structure

	Attribute 1	Attribute 2	Attribute 3
Case 1			
Case 2			
Case 3			
Case 4			

Data Structure

Attributes

Cases



```
{r}
head(lego)
```

A tibble: 6 × 14

item_number <dbl>	set_name <chr>	theme <chr>	pieces <dbl>	price <dbl>
10859	My First Ladyb…	DUPLO®	6	4.99
10860	My First Race …	DUPLO®	6	4.99
10862	My First Celeb…	DUPLO®	41	14.99
10864	Large Playgro…	DUPLO®	71	49.99
10867	Farmers' Market	DUPLO®	26	19.99
10870	Farm Animals	DUPLO®	16	9.99

6 rows | 1–5 of 14 columns

Variable Types

- Categorical
  - Recorded as words (or categories/levels)
  - Example: Highest Degree (High School, Bachelors, etc)
  - Quantitative
    - Recorded as a number
    - Example: Height (recorded in inches)

Type: Categorical

- "theme" is a categorical attribute
- The first 6 in our data set are all DUPLO themed

```
{r}
head(lego)
```

A tibble: 6 × 14

item_number <dbl>	set_name <chr>	theme <chr>	pieces <dbl>	price <dbl>
10859	My First Ladyb...	DUPLO®	6	4.99
10860	My First Race ...	DUPLO®	6	4.99
10862	My First Celeb...	DUPLO®	41	14.99
10864	Large Playgro...	DUPLO®	71	49.99
10867	Farmers' Market	DUPLO®	26	19.99
10870	Farm Animals	DUPLO®	16	9.99

6 rows | 1-5 of 14 columns

Type: Quantitative

- "pieces" is a quantitative attribute

```
{r}
head(lego)
```

A tibble: 6 × 14

item_number <dbl>	set_name <chr>	theme <chr>	pieces <dbl>	price <dbl>
10859	My First Ladyb…	DUPLO®	6	4.99
10860	My First Race …	DUPLO®	6	4.99
10862	My First Celeb…	DUPLO®	41	14.99
10864	Large Playgro…	DUPLO®	71	49.99
10867	Farmers' Market	DUPLO®	26	19.99
10870	Farm Animals	DUPLO®	16	9.99

6 rows | 1–5 of 14 columns

GOAL: Learn about our attributes

```
{r}
head(lego)
```

A tibble: 6 × 14

item_number <dbl>	set_name <chr>	theme <chr>	pieces <dbl>	price <dbl>
10859	My First Ladyb…	DUPLO®	6	4.99
10860	My First Race …	DUPLO®	6	4.99
10862	My First Celeb…	DUPLO®	41	14.99
10864	Large Playgro…	DUPLO®	71	49.99
10867	Farmers' Market	DUPLO®	26	19.99
10870	Farm Animals	DUPLO®	16	9.99

6 rows | 1–5 of 14 columns

Datasets themselves are not useful for this

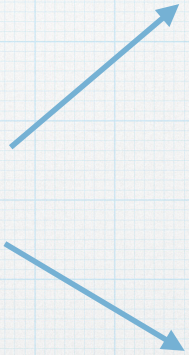
GOAL: Learn about our attributes

```r
head(lego)
```

A tibble: 6 × 14

item_number <dbl>	set_name <chr>	theme <chr>	pieces <dbl>	price <dbl>
10859	My First Ladyb…	DUPLO®	6	4.99
10860	My First Race …	DUPLO®	6	4.99
10862	My First Celeb…	DUPLO®	41	14.99
10864	Large Playgro…	DUPLO®	71	49.99
10867	Farmers' Market	DUPLO®	26	19.99
10870	Farm Animals	DUPLO®	16	9.99

6 rows | 1–5 of 14 columns



Statistics

Graphs

Categorical Data

Summary Statistics for Categorical

- Counts
- Percents
- Proportions

```r
df_stats( ~ theme, data = lego, "counts")
```

Description: df [1 × 4]

response <chr>	n_City <int>	n_DUPLO® <int>	n_Friends <int>
theme	25	25	25

1 row

```r
df_stats( ~ theme, data = lego, "props")
```

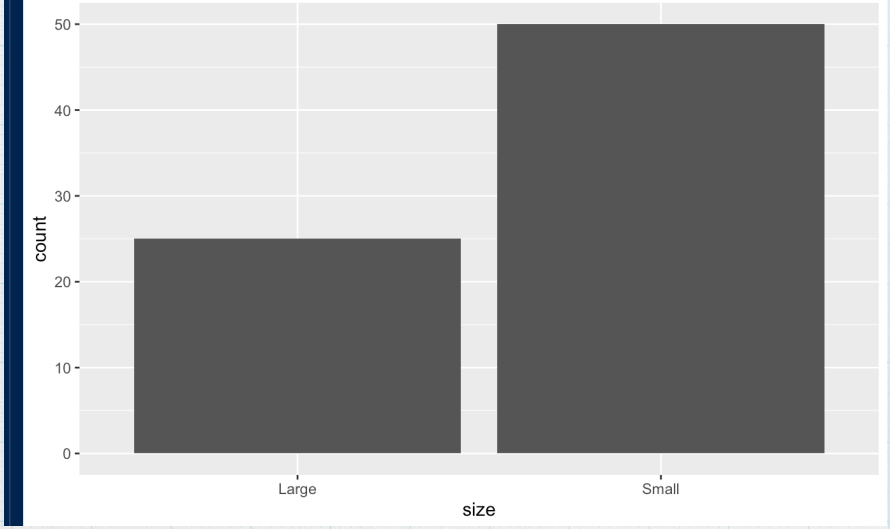
Description: df [1 × 4]

response <chr>	prop_City <dbl>	prop_DUPLO® <dbl>	prop_Friends <dbl>
theme	0.3333333	0.3333333	0.3333333

1 row

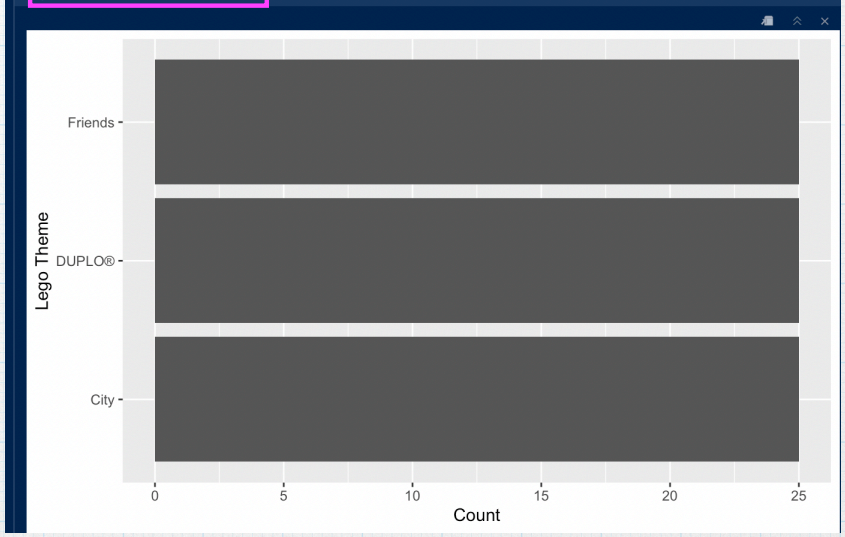
Graphs for Categorical Data

```r
gf_counts(
  ~size, data = lego
  )
```



Vertical Bar Chart for Size Box

```r
gf_countsh(
  ~theme, data = lego,
  xlab = "Count",
  ylab = "Lego Theme"
  )
```



Horizontal Bar Chart for Theme

Other Graphs for Categorical Data

- Pie Charts
- Donut Charts

Quantitative Data

Summary Statistics for Quantitative Data

- Mean
- Median
- Mode

```r
df_stats(~price, data = lego)
```

Description: df [1 × 10]

respo... <chr>	min <dbl>	Q1 <dbl>	median <dbl>	Q3 <dbl>	max <dbl>	mean <dbl>	sd <dbl>	n <int>
price	4.99	9.99	19.99	39.99	149.99	32.13...	27.74...	75

1 row | 1–9 of 10 columns

```r
df_stats(~pieces, data = lego)
```

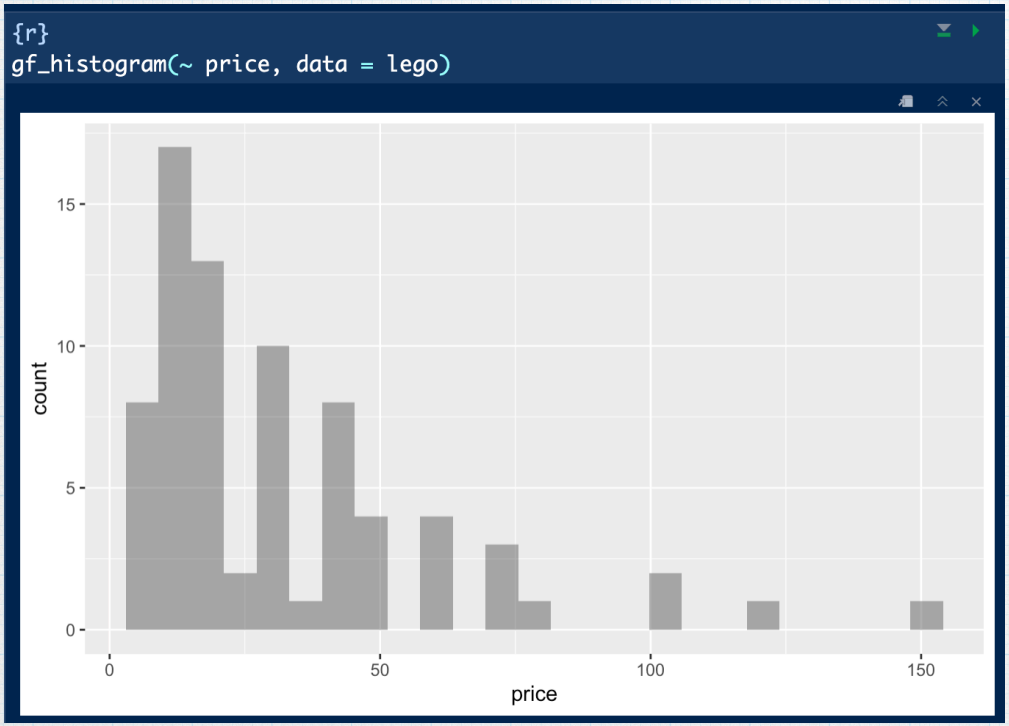
Description: df [1 × 10]

respo... <chr>	min <dbl>	Q1 <dbl>	median <dbl>	Q3 <dbl>	max <dbl>	mean <dbl>	sd <dbl>	n <int>
pieces	6	37	93	323.5	897	196.4...	213.9...	75

1 row | 1–9 of 10 columns

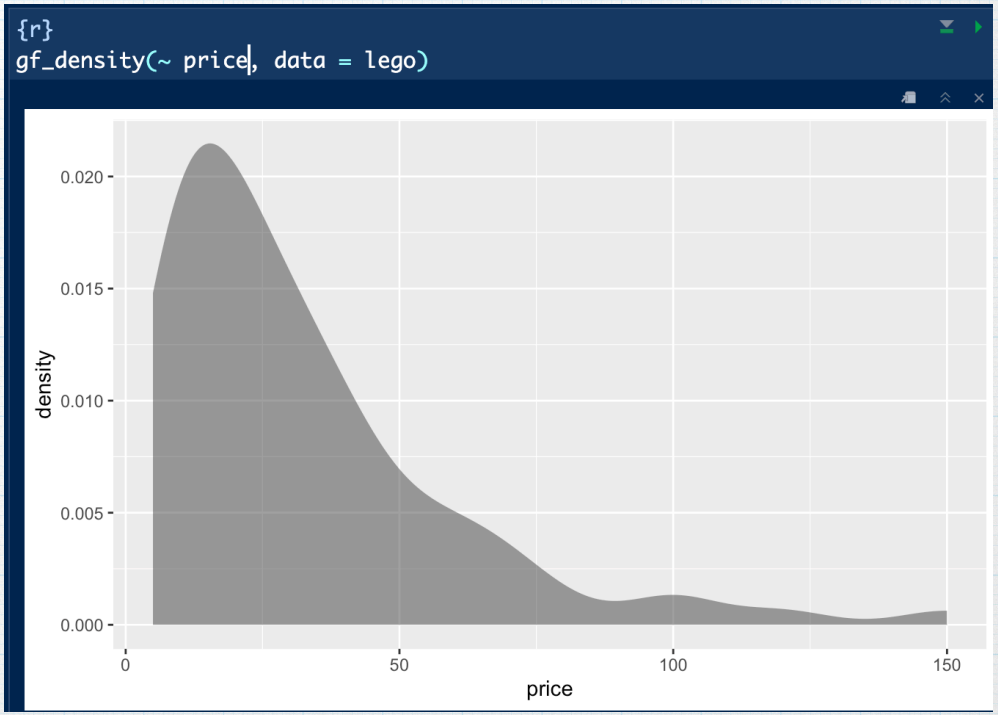
Graphs for Quantitative Data

```r
gf_histogram(~ price, data = lego)
```



Histogram of Price

```r
gf_density(~ price, data = lego)
```



Density Plot of Price

Other Graphs for Quantitative Data

- Boxplots
- Dotplots
- Density Plots

Variables and Graphs Activity

Summary

- Data is made up of cases (rows) and attributes (columns)
- Our attributes are either categorical (a word) or quantitative (a number)
- We can create graphs for each variable type
- We can create summary statistics for each variable type