Introduction to the Grammar of Graphics III

Nate Wells

Math 141, 2/3/21

Outline

In this lecture, we will. . .

Outline

In this lecture, we will. . .

- Discuss Linegraphs, Histrograms, Boxplots, and Barplots
- Investigate some options for further customizing graphs

Section 1

Common Graphs using `ggplot2`

The Five Named Graphs

- We focus on just 5 graphs fundamental to statistics (although other types exist)

The Five Named Graphs

- We focus on just 5 graphs fundamental to statistics (although other types exist)
  1. Scatterplots
  2. **Linegraphs**
  3. **Histograms**
  4. **Boxplots**
  5. **Barplots**

The Five Named Graphs

- We focus on just 5 graphs fundamental to statistics (although other types exist)
    1. Scatterplots
    2. **Linegraphs**
    3. **Histograms**
    4. **Boxplots**
    5. **Barplots**
- We'll use a common data set to investigate each graph: the Portland Biketown data

```
biketown <-
  read_csv("biketown.csv")
```

Line Graphs

How do bike use patterns change throughout the day?

Line Graphs

How do bike use patterns change throughout the day?

```
biketown2 <- count(biketown, StartHour)
biketown2
```

```
## # A tibble: 24 x 2
##    StartHour     n
##        <int> <int>
##  1         0   118
##  2         1    69
##  3         2    50
##  4         3    20
##  5         4    35
##  6         5    71
##  7         6   104
##  8         7   270
##  9         8   492
## 10         9   392
## # ... with 14 more rows
```

Line Graphs

- Frequently, we compare two quantitative variables where one variable represents time. It is illustrative to connect neighboring points with a smooth curve.

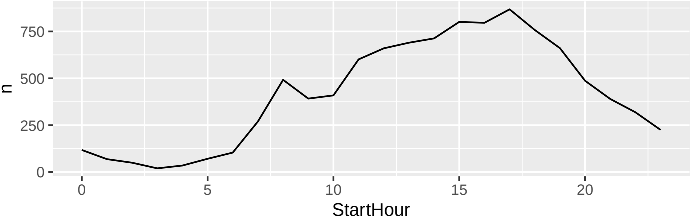
Line Graphs

- Frequently, we compare two quantitative variables where one variable represents time. It is illustrative to connect neighboring points with a smooth curve.
- These **line graphs** (or time series) provide stronger sequential and/or cyclic visual cues.

Line Graphs

- Frequently, we compare two quantitative variables where one variable represents time. It is illustrative to connect neighboring points with a smooth curve.
- These **line graphs** (or time series) provide stronger sequential and/or cyclic visual cues.

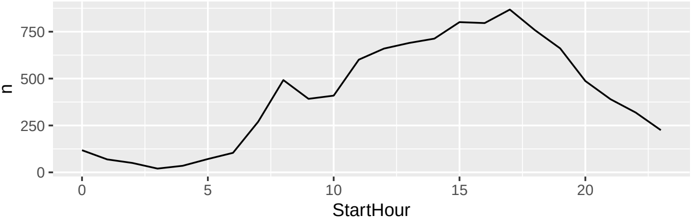
```
ggplot(data = biketown2, mapping = aes(x = StartHour, y = n)) +
  geom_line()
```



Line Graphs

- Frequently, we compare two quantitative variables where one variable represents time. It is illustrative to connect neighboring points with a smooth curve.
- These **line graphs** (or time series) provide stronger sequential and/or cyclic visual cues.

```
ggplot(data = biketown2, mapping = aes(x = StartHour, y = n)) +
  geom_line()
```



- To construct a line graph , use `geom_line()` with the aesthetic mapping `aes(x = ... , y = ...)`.

The Distribution of a Variable

- Consider the Distance variable in the biketown data set. What are its minimum, maximum, and central values?

The Distribution of a Variable

- Consider the `Distance` variable in the `biketown` data set. What are its minimum, maximum, and central values?
- What proportion of observations are “close” to these extremes?

The Distribution of a Variable

- Consider the `Distance` variable in the `biketown` data set. What are its minimum, maximum, and central values?
- What proportion of observations are “close” to these extremes?
- These questions can be answered by exploring the **distribution** of a variable, which is a representation of the unique values it takes along with the frequency it takes them.

Histograms

- Distributions are most commonly visualized by way of the **histogram**

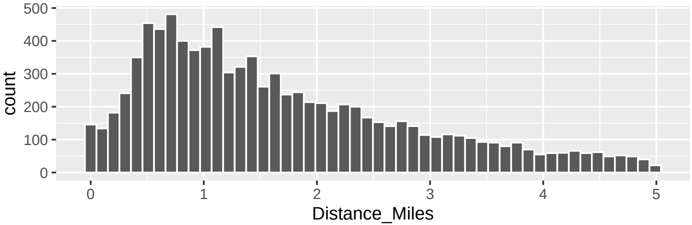
Histograms

- Distributions are most commonly visualized by way of the **histogram**
- To create a histogram:
  - Divide the x-axis into a sequence of equally-sized intervals (or bins).
  - For each, count the number of observations falling in that interval.
  - Draw bars with height equal to count and with width spanning the interval.

Histograms

- Distributions are most commonly visualized by way of the **histogram**
- To create a histogram:
  - Divide the x-axis into a sequence of equally-sized intervals (or bins).
  - For each, count the number of observations falling in that interval.
  - Draw bars with height equal to count and with width spanning the interval.

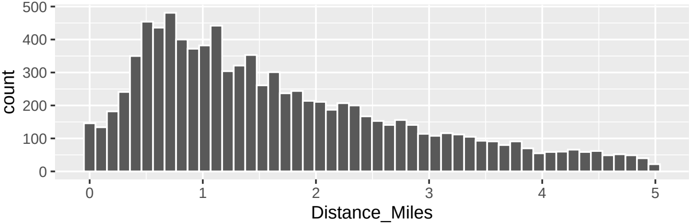
```
ggplot(data = biketown_short, mapping = aes(x = Distance_Miles)) +
  geom_histogram(bins = 50, color = "White")
```



Histograms

- Distributions are most commonly visualized by way of the **histogram**
- To create a histogram:
  - Divide the x-axis into a sequence of equally-sized intervals (or bins).
  - For each, count the number of observations falling in that interval.
  - Draw bars with height equal to count and with width spanning the interval.

```
ggplot(data = biketown_short, mapping = aes(x = Distance_Miles)) +
  geom_histogram(bins = 50, color = "White")
```



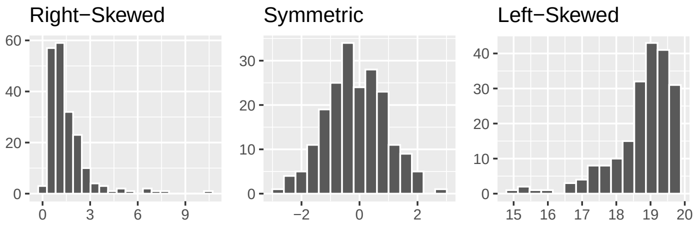
- Minimimum? Maximum? Center? Spread?

The Shape of You (Distributions)

- Histograms also reveal qualitative information about the shape of a variable's distribution:

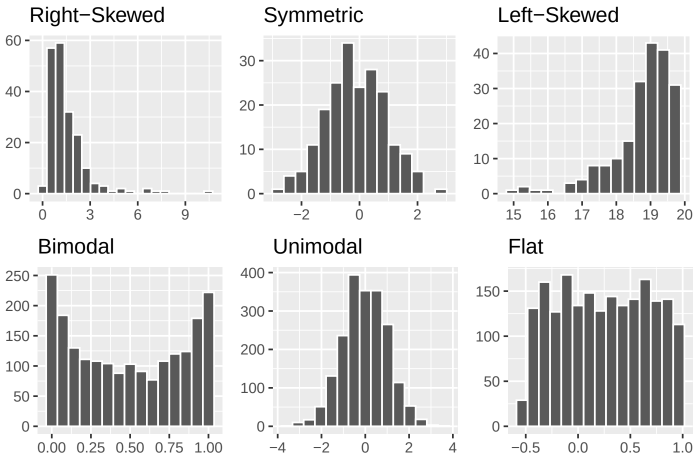
The Shape of You (Distributions)

- Histograms also reveal qualitative information about the shape of a variable's distribution:



The Shape of You (Distributions)

- Histograms also reveal qualitative information about the shape of a variable's distribution:



How many bins?

- The number of bins used can radically affect the shape of the histogram.

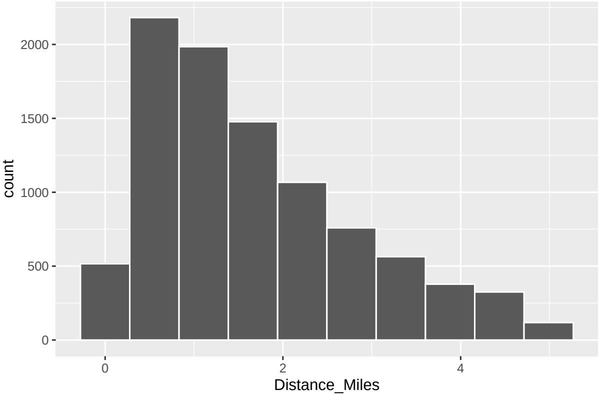
How many bins?

- The number of bins used can radically affect the shape of the histogram.
    - Use `bins=` to set the number of bins in a histogram

How many bins?

- The number of bins used can radically affect the shape of the histogram.
  - Use bins= to set the number of bins in a histogram

```
ggplot(data = biketown_short, mapping = aes(x = Distance_Miles))+
  geom_histogram(bins=10, color = "white")
```

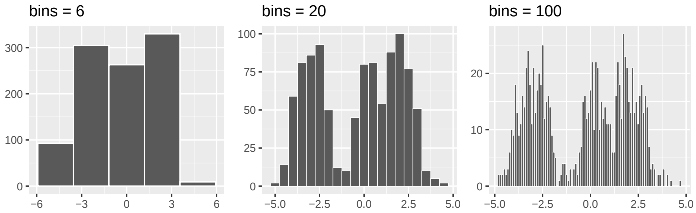


The Effect of Bin Size

- Each of the following is a histogram for *the same data*, with different values for the `bins =` argument in `geom_histogram()`

The Effect of Bin Size

- Each of the following is a histogram for *the same data*, with different values for the `bins =` argument in `geom_histogram()`



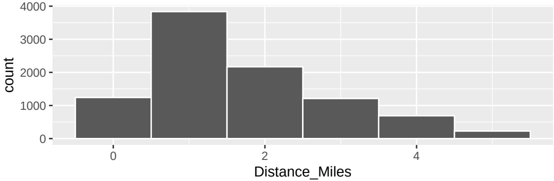
How many bins?

- Alternatively, we can specify the width of bins using `binwidth =`

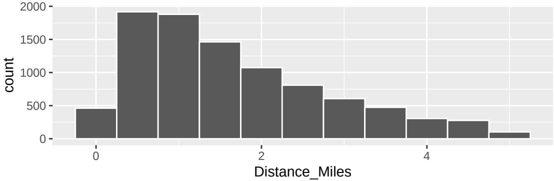
How many bins?

- Alternatively, we can specify the width of bins using `binwidth =`

```
ggplot(data = biketown_short, mapping = aes(x = Distance_Miles))+
  geom_histogram(binwidth = 1, color = "white")
```



```
ggplot(data = biketown_short, mapping = aes(x = Distance_Miles))+
  geom_histogram(binwidth = 0.5, color = "white")
```



Summary Statistics

- The **five-number summary** of a data set consists of: Minimum, 1st Quartile ($Q1$), Median, 3rd Quartile ($Q3$), Maximum.

Summary Statistics

- The **five-number summary** of a data set consists of: Minimum, 1st Quartile ($Q1$), Median, 3rd Quartile ($Q3$), Maximum.
- The **median** is a value so that 50% of data lies above it and 50% lies below.

Summary Statistics

- The **five-number summary** of a data set consists of: Minimum, 1st Quartile ($Q1$), Median, 3rd Quartile ($Q3$), Maximum.
- The **median** is a value so that 50% of data lies above it and 50% lies below.
- The **1st / 3rd quartiles** are values so that 25% / 75% of data lies below it and 75% / 25% lies above.

Summary Statistics

- The **five-number summary** of a data set consists of: Minimum, 1st Quartile ($Q1$), Median, 3rd Quartile ($Q3$), Maximum.
- The **median** is a value so that 50% of data lies above it and 50% lies below.
- The **1st / 3rd quartiles** are values so that 25% / 75% of data lies below it and 75% / 25% lies above.
- The median separates the data into two equal parts. Note $Q1$ is also the median of the lower part, while $Q3$ is the median of the upper part.

Summary Statistics

- The **five-number summary** of a data set consists of: Minimum, 1st Quartile ($Q1$), Median, 3rd Quartile ($Q3$), Maximum.
- The **median** is a value so that 50% of data lies above it and 50% lies below.
- The **1st / 3rd quartiles** are values so that 25% / 75% of data lies below it and 75% / 25% lies above.
- The median separates the data into two equal parts. Note $Q1$ is also the median of the lower part, while $Q3$ is the median of the upper part.
- The **interquartile range** (IQR) is $Q3 - Q1$ and measures the spread of the middle 50% of the data.

Summary Statistics

- The **five-number summary** of a data set consists of: Minimum, 1st Quartile ($Q1$), Median, 3rd Quartile ($Q3$), Maximum.
- The **median** is a value so that 50% of data lies above it and 50% lies below.
- The **1st / 3rd quartiles** are values so that 25% / 75% of data lies below it and 75% / 25% lies above.
- The median separates the data into two equal parts. Note $Q1$ is also the median of the lower part, while $Q3$ is the median of the upper part.
- The **interquartile range** (IQR) is $Q3 - Q1$ and measures the spread of the middle 50% of the data.
- Taken together, the five-number summary provides a measure of center and spread of a data set.

Boxplots

- The five-number summary can be visualized by way of the boxplot.

Boxplots

- The five-number summary can be visualized by way of the boxplot.
- Consider the five number summary for `Distance_miles` in the biketown data

Boxplots

- The five-number summary can be visualized by way of the boxplot.
- Consider the five number summary for `Distance_miles` in the biketown data

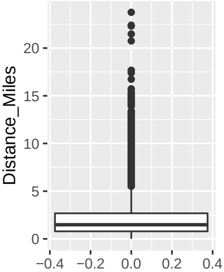
Min	Q1	Median	Q3	Max
0	0.79	1.48	2.68	23.75

Boxplots

- The five-number summary can be visualized by way of the boxplot.
- Consider the five number summary for Distance_miles in the biketown data

Min	Q1	Median	Q3	Max
0	0.79	1.48	2.68	23.75

```r
ggplot(data=biketown,
       mapping=aes(y=Distance_Miles))+
  geom_boxplot()
```

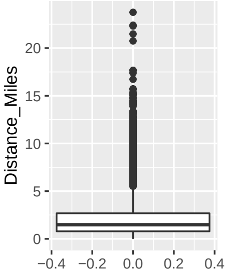


Boxplots

- The five-number summary can be visualized by way of the boxplot.
- Consider the five number summary for `Distance_miles` in the biketown data

Min	Q1	Median	Q3	Max
0	0.79	1.48	2.68	23.75

```r
ggplot(data=biketown,
       mapping=aes(y=Distance_Miles))+
  geom_boxplot()
```



- The Top / Bottom of box correspond to $Q3$ / $Q1$, while center line is median.
- The "whiskers" extend $1.5 \times \mathrm{IQR}$ in either direction from box edge.
- **Outliers** are any observations outside this range, and are plotted as points.

Side-by-side Boxplots

- Often, we compare the distribution of a variable conditioned on values of a 2nd.

Side-by-side Boxplots

- Often, we compare the distribution of a variable conditioned on values of a 2nd.
- To do so, include an x-position aesthetic mapping from the 2nd variable.

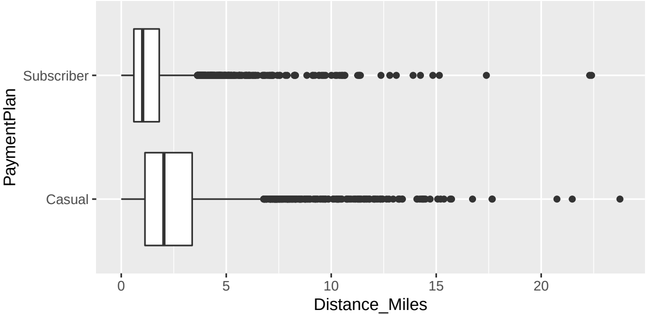
Side-by-side Boxplots

- Often, we compare the distribution of a variable conditioned on values of a 2nd.
- To do so, include an x-position aesthetic mapping from the 2nd variable.
- To have boxes span horizontally, rather than vertically, add a `coord_flip()` layer.

Side-by-side Boxplots

- Often, we compare the distribution of a variable conditioned on values of a 2nd.
- To do so, include an *x*-position aesthetic mapping from the 2nd variable.
- To have boxes span horizontally, rather than vertically, add a `coord_flip()` layer.

```
ggplot(data = biketown, mapping = aes(x = PaymentPlan, y = Distance_Miles)) +
  geom_boxplot()+ coord_flip()
```



Bar Charts

- Both Boxplots and Histograms show the distribution of *quantitative* variables.

Bar Charts

- Both Boxplots and Histograms show the distribution of *quantitative* variables.
- We use Bar Charts to visualize the distribution of *categorical* variables, whose values are broken down into distinct levels.

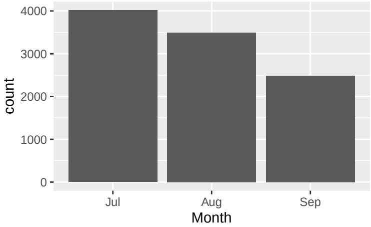
Bar Charts

- Both Boxplots and Histograms show the distribution of *quantitative* variables.
- We use Bar Charts to visualize the distribution of *categorical* variables, whose values are broken down into distinct levels.
- Investigate the distribution of bike use by `month`

Bar Charts

- Both Boxplots and Histograms show the distribution of *quantitative* variables.
- We use Bar Charts to visualize the distribution of *categorical* variables, whose values are broken down into distinct levels.
- Investigate the distribution of bike use by `month`

```
ggplot(data = biketown, mapping = aes(x = Month)) +
  geom_bar()
```



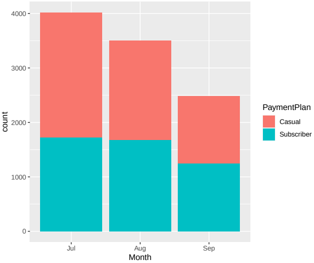
Segmented / Stacked Bar Charts

- Bar charts used to visualize the *joint distribution* of 2 categorical variables.

Segmented / Stacked Bar Charts

- Bar charts used to visualize the *joint distribution* of 2 categorical variables.

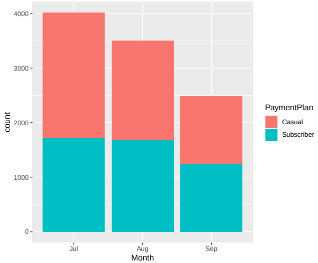
```
ggplot(data = biketown,
       mapping = aes(x = Month,
                     fill = PaymentPlan)) +
  geom_bar()
```



Segmented / Stacked Bar Charts

- Bar charts used to visualize the *joint distribution* of 2 categorical variables.

```
ggplot(data = biketown,
       mapping = aes(x = Month,
                    fill = PaymentPlan)) +
  geom_bar()
```

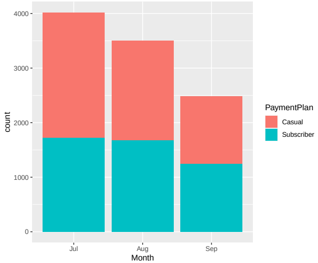


- Each bar divided into count by `fill` variable.

Segmented / Stacked Bar Charts

- Bar charts used to visualize the *joint distribution* of 2 categorical variables.

```
ggplot(data = biketown,
       mapping = aes(x = Month,
                     fill = PaymentPlan)) +
  geom_bar()
```

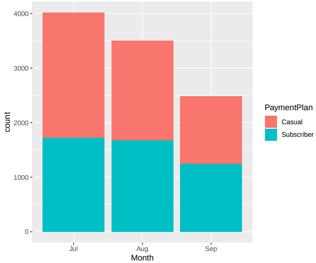


- Each bar divided into count by `fill` variable.
- Hard to make direct comparisons

Segmented / Stacked Bar Charts

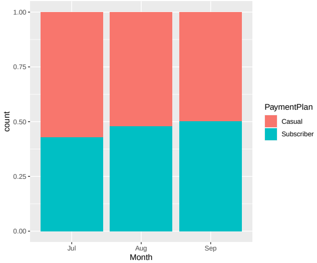
- Bar charts used to visualize the *joint distribution* of 2 categorical variables.

```r
ggplot(data = biketown,
       mapping = aes(x = Month,
                     fill = PaymentPlan))
  geom_bar()
```



- Each bar divided into count by `fill` variable.
- Hard to make direct comparisons

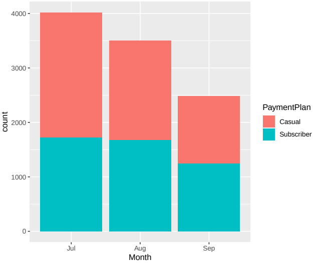
```r
ggplot(data = biketown,
       mapping = aes(x = Month,
                     fill = PaymentPlan))
  geom_bar(position = "fill")
```



Segmented / Stacked Bar Charts

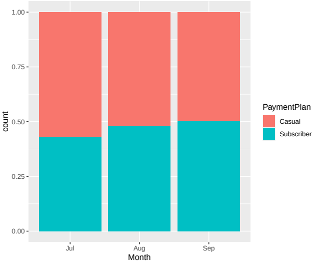
- Bar charts used to visualize the *joint distribution* of 2 categorical variables.

```
ggplot(data = biketown,
       mapping = aes(x = Month,
                     fill = PaymentPlan))
  geom_bar()
```



- Each bar divided into count by `fill` variable.
- Hard to make direct comparisons

```
ggplot(data = biketown,
       mapping = aes(x = Month,
                     fill = PaymentPlan))
  geom_bar(position = "fill")
```



- Each bar divided into proportion by `fill` variable.

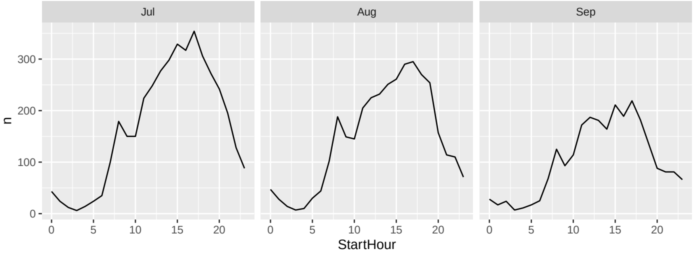
Facets

- Faceting is used to split one graphic into many smaller ones, based on the values of a categorical variable.

Facets

- Faceting is used to split one graphic into many smaller ones, based on the values of a categorical variable.

```
ggplot(data = biketown2, mapping = aes(x = StartHour, y = n)) +
  geom_line() +
  facet_wrap(~Month, ncol = 3)
```

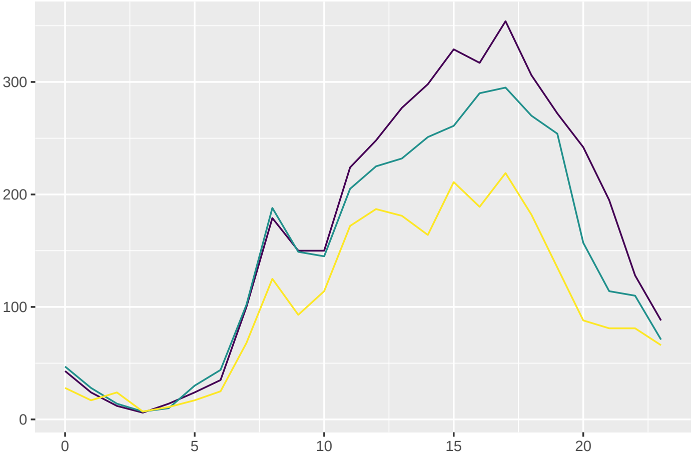


Adding Context

- Adding titles and axes labels to graphs greatly improves clarity.

Adding Context

- Adding titles and axes labels to graphs greatly improves clarity.



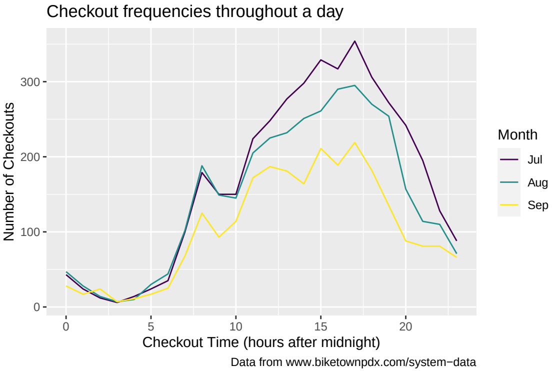
Adding Context

- Adding titles, captions, and axis labels greatly improves clarity.

Adding Context

- Adding titles, captions, and axis labels greatly improves clarity.

```r
ggplot(data = biketown2, mapping = aes(x = StartHour, y = n, color = Month)) +
  geom_line( ) +
  labs(x = "Checkout Time (hours after midnight)", y = "Number of Checkouts",
       title  = "Checkout frequencies throughout a day",
       caption = "Data from www.biketownpdx.com/system-data")
```


Checkout frequencies throughout a day
Number of Checkouts
Checkout Time (hours after midnight)
Month
Jul
Aug
Sep
Data from www.biketownpdx.com/system−data