Visualisierung mit Grafiken

Um uns einen Überblick über die Daten zu verschaffen und einfacher Zusammenhänge zwischen Variablen zu erhalten, ist es oft sinnvoll, Diagramme zu erstellen. R bietet dafür zahlreiche Möglichkeiten, die sehr flexibel genutzt werden können.

In dieser Einheit werden wir mit dem Datensatz `m111survey` aus dem `tigerstats`-Paket die Grundlagen für eine solche Visualisierung erarbeiten. Dieser Datensatz ist das Ergebnis einer Umfrage unter 71 Studierenden des Georgetown College in Kentucky.

Zunächst laden wir diesen Datensatz in unsere Umgebung.

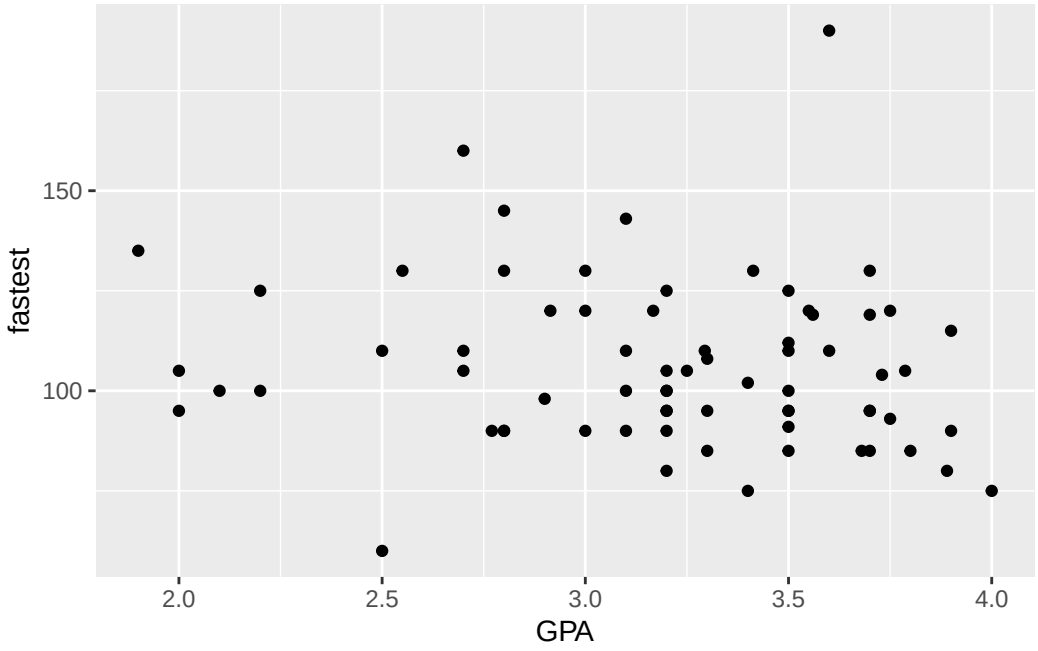
```
install.packages("tigerstats")
```

```
library(tigerstats)
```

Mithilfe von `?m111survey` können wir uns ansehen, welche Bedeutung die Variablen haben. Man könnte zum Beispiel annehmen, dass Studenten mit niedrigerem Notendurchschnitt eher dazu neigen, schnell zu fahren. Veranschaulichen wir diesen Zusammenhang in einem Diagramm:

```
m111survey %>%
  ggplot(aes(x = GPA, y = fastest)) +
  geom_point()
```

```
Warning: Removed 1 rows containing missing values (`geom_point()`).
```



Dieses Diagramm ist ein sogenannter *Scatterplot* bzw. ein Streudiagramm. Der Aufbau eines solchen Diagramms in R ist wie folgt:

```
datensatz %>%
  # Legt den "Rahmen" und die Achsen fest
  ggplot(aes(x = x-Achsen-Variable, y = y-Achsen-Variable)) +
  geom_... + # Das sogenannte "Layer", z.B. Punkte, Linien, Formen, ...
  ... # Man kann mehrere Layer übereinander legen.
```

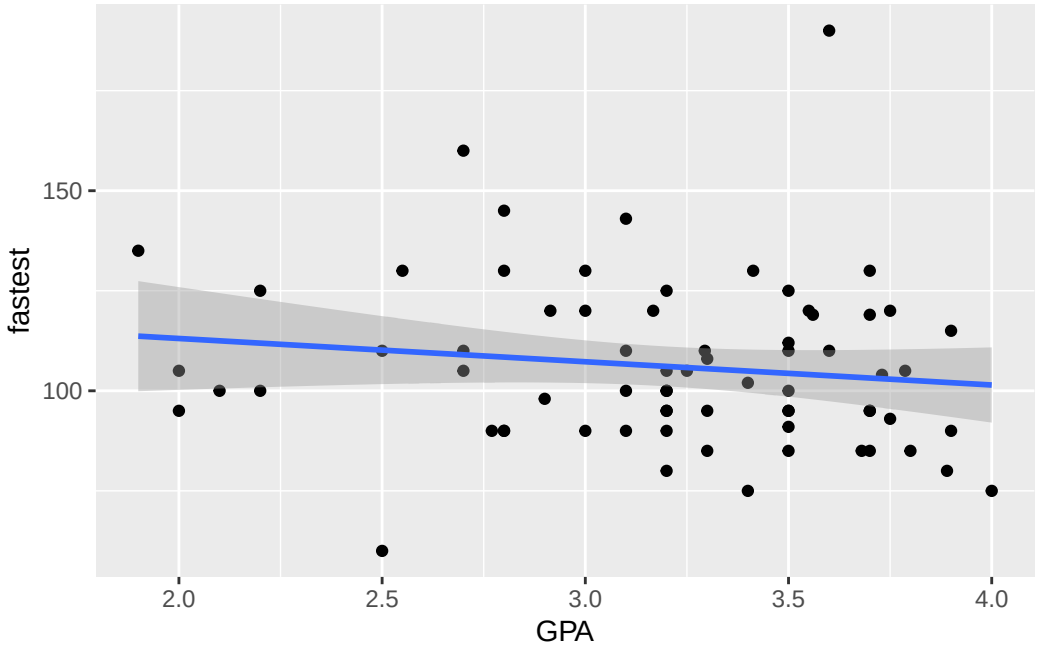
Man könnte denken, dass diese Annahme tatsächlich stimmt. Legen wir mal eine "Ausgleichsgerade" durch die Punkte:

```
m111survey %>%
  ggplot(aes(x = GPA, y = fastest)) +
  geom_point() +
  geom_smooth(method = "lm") # lm steht für "Linear Model"
```

```
`geom_smooth()` using formula = 'y ~ x'
```

```
Warning: Removed 1 rows containing non-finite values (`stat_smooth()`).
```

```
Warning: Removed 1 rows containing missing values (`geom_point()`).
```



Der graue Balken in diesem Diagramm beschreibt im Grunde die Unsicherheit in unserem Modell. Wir sehen, dass er so breit ist, dass die Enden sich jeweils überlappen. Daher können wir nicht darauf schließen, dass der Notenschnitt einen Einfluss auf die höchste je gefahrene Geschwindigkeit hat.

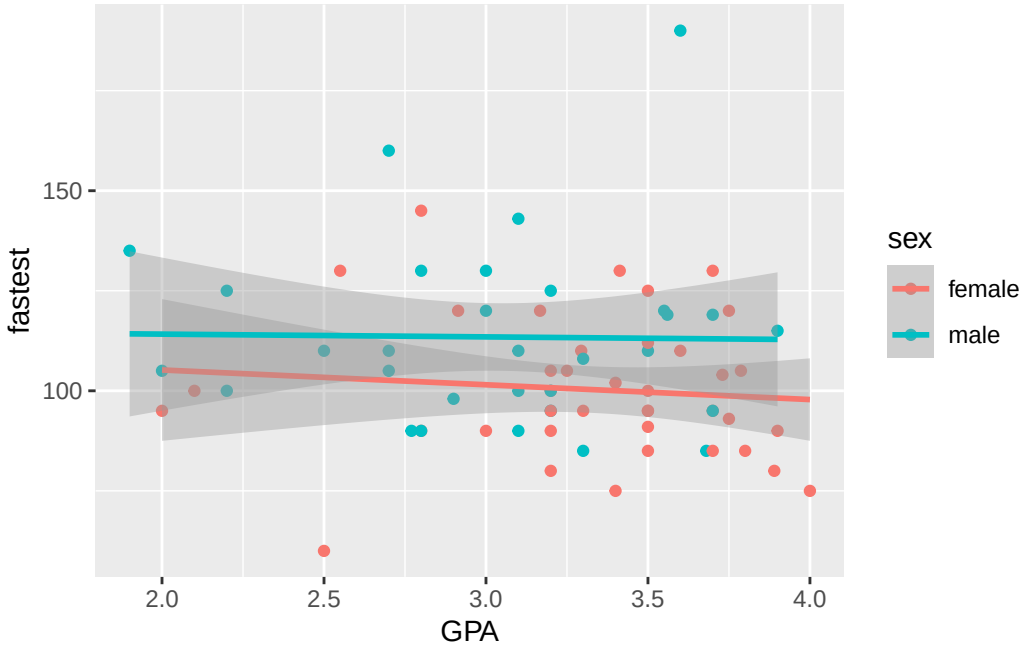
Wir können aber mehr aus den Daten holen. Schauen wir uns zum Beispiel an, wie Frauen im Vergleich zu Männern in der Umfrage antworten:

```
m111survey %>%
  ggplot(aes(x = GPA, y = fastest, color = sex)) +
  geom_point() +
  geom_smooth(method = "lm")
```

```
`geom_smooth()` using formula = 'y ~ x'
```

```
Warning: Removed 1 rows containing non-finite values (`stat_smooth()`).
```

```
Warning: Removed 1 rows containing missing values (`geom_point()`).
```

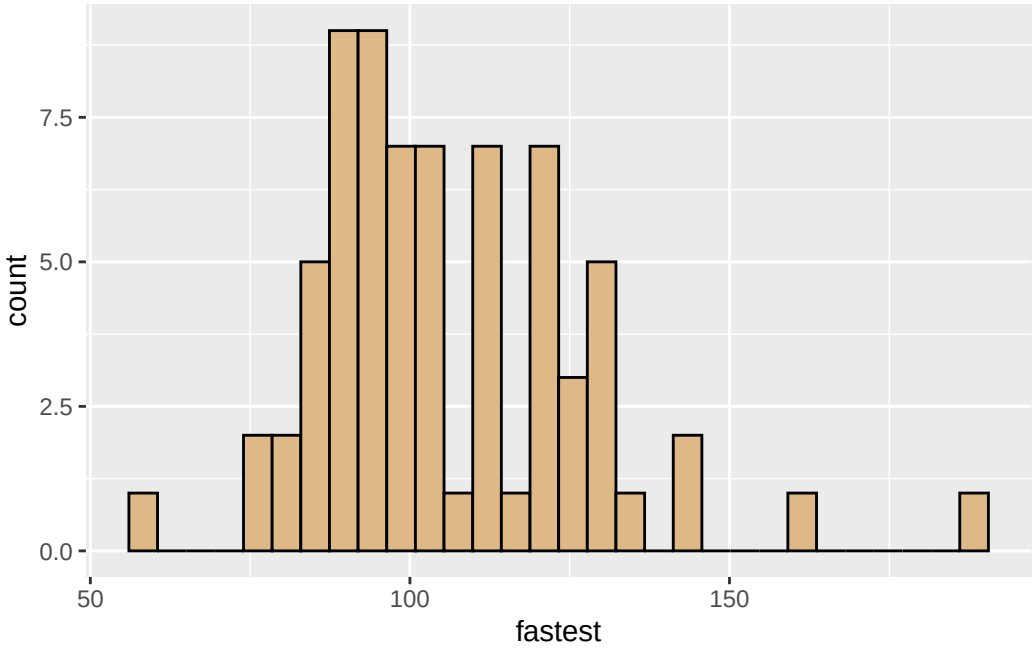


Wir sehen: Frauen haben tendenziell langsamere Höchstgeschwindigkeiten als Männer, aber auch hier überlappen sich die Balken.

Ein weiterer wichtiger Diagrammtyp ist das *Histogramm*. Es gibt an, wie oft ein bestimmtes "Level" einer Variable in den Daten vorkommt, zum Beispiel wie folgt:

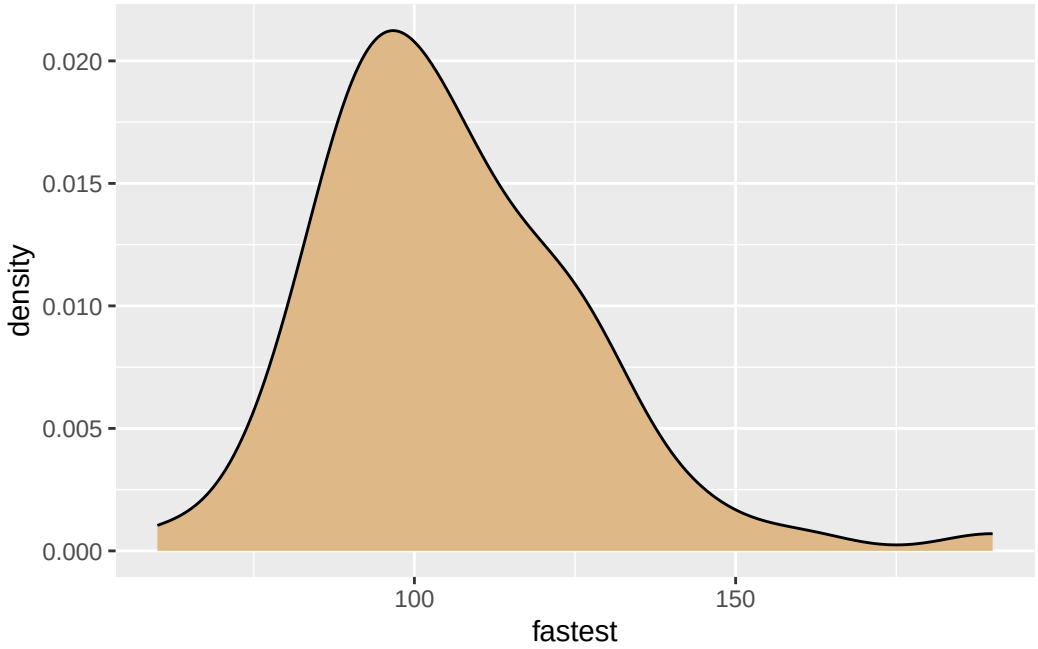
```
m111survey %>%
  ggplot(aes(x = fastest)) +
  geom_histogram(fill = "burlywood", color = "black")
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Ähnliches kann durch das Layer `geom_density()` erreicht werden. Dieses gibt nicht die Anzahl, sondern die Verteilung der Geschwindigkeiten an.

```
m111survey %>%
  ggplot(aes(x = fastest)) +
  geom_density(fill = "burlywood")
```


0.020
0.015
0.010
0.005
0.000
density
100
150
fastest