3 Data visualisation

**Grundlagen mit ggplot**

* ggplot2 und der Befehlsaufbau
  + Aufbau:
    - ggplot(data = <DATA>) + #Datensatz der Grafik definieren
      * <GEOM\_FUNCTION>( #Grafiklayer auf Basis der Plotdaten
        + mapping = aes(<MAPPINGS>),
        + stat = <STAT>,
        + position = <POSITION>
        + ) +
      * <COORDINATE\_FUNCTION> +
      * <FACET\_FUNCTION>
  + Beispiel:
    - ggplot(tibble) +
      * geom\_point(mapping=(aes(x=var1, y=var2))
  + Cheatsheet: <http://rstudio.com/resources/cheatsheets>
* Geommetrische Figuren mit Geom\_Function
  + graphische Beschreibung der Daten
  + \_point, \_smooth, etc.: <https://exts.ggplot2.tidyverse.org/gallery/>
  + Kombinierbar mit aes und facets
* Anpassung des Mappings durch **aes**, kurz: Ästhetik
  + Farblich mit aes(x=x, y=y, **color=category)**
  + Größe mit aes(x=x, y=y, **size=category)**
  + Transparenz mit aes(x=x, y=y, **alpha=category)**
  + Form mit aes(x=x, y=y, **shape=category)**
  + **Kann in ggplot oder in geom untergebracht werden**
    - In Ggplot wirkt es als globale Variable
    - In geom überschreibt es für das Layer die Variablen
* **Facets = Gruppierende Anordnung nach kategorialen Variablen**
  + facet\_wrap(~ class, nrow = 2)
    - erzeugt Grafiken nach Kategorie sortiert auf 2 Zeilen
    - variablen müssen diskret sein
  + facet\_grid(drv ~ cyl)
    - erzeugt Grafiken nach der Kombination zweier Kategorien gruppiert
    - mit (. ~ cyl) oder (cyl ~ .) ändert sich die gruppierung mit einer Kategorie entweder auf Zeilen oder Spalten Anordnung

**Statistische Transformation**

* Balkendiagramme, Histogramme und Häufigkeitspolygone binden Daten ein und stellen dann die Häufigkeitsrepräsentation in den definierten Gruppen dar
* Smoother passen ein Modell an die Daten an und stellen dann die Vorhersagen des Modells dar
* Boxplots berechnen eine robuste Zusammenfassung der Verteilung und zeigen dann eine speziell formatierte Box

Ablauf im Hintergrund:

* Beim Aufruf der geom\_function wird stat() ausgeführt und berechnet für den Anwendungsfall die statistischen Parameter, die im jeweiligen Plot gewünscht sind
  + z.B. ruft geom\_bar() die Funktion stat\_count() auf, um die absolute Häufigkeit der definierten Gruppen zu berechnen
  + geom und stat sind miteinander gekoppelt 🡪 der Aufruf eines geom nutzt automatisch ein statistisches Verfahren und umgekehrt. Daher ist die Zeichnung eines grafen auch mit der statistischen Funktion möglich.
* Auflistung der statistischen Funktionen unter ?stat\_bin

**Positionsanpassung**

* Argument: position =
  + „identity“:
  + „dodge“: zeigt sich eigentlich überlappende Gruppen nebeneinander an
  + „fill“: zeigt die Proportion der Verteilung innerhalb einer Gruppe
  + Jitter: für Punktdiagramme, zieht die Punkte etwas auseinander um die Dichte von verteilten Punkten anzuzeigen
* Hilfe über: ?position\_dodge, ?position\_fill, ?position\_identity, ?position\_jitter, and ?position\_stack

**Koordinatensystem**

* Methoden:
  + Cord\_flip: Tausch x- und y-Achse
  + Cord\_quickmap: passt die Grafik and das Verhältnis an
  + Cord\_polar: CoxComb-Chart aus einem Barchart wird erzeugt

**Kapitel 7: Exploratory Data Analysis / Explorative Datenanalyse**

EDA-Phasen:

* Fragen zu den Daten generieren
* Antworten über Visualisierung, Transformierung und Modellierung der Daten
* Nutze Erkenntnisse um die Fragen und Darstellung anzupassen

Ziel:

* Am Anfang: Qualitative Fragen stellen um quantitative Fragen zu generieren

Verallgemeinerte Fragen:

* What type of variation occurs within my variables?
* What type of covariation occurs between my variables?

Grundbegriffe:

* Variable: eine messbare Quantität, Qualität oder Eigenschaft
* Value/Wert: Zustand einer Variable/ Ausprägung
* Oberservation/Beobachtung: Set an Messungen unter derselben Bedingung (örtlich, zeitlich)
* Tabelarische Daten: Eine Sammlung von Werten, die in Variablen und Beobachtungen angeordnet sind

**Variation:**

* Tendenz das Werte einer Variable sich von Messpunkt zu Messpunkt unterscheiden
* Darstellung über:
  + Verteilung
    - Balkendiagram oder Histogramm (nominal oder metrisch)
      * Häufig/Selten anzutreffende Werte?
      * Werden Erwartungen erfüllt/ Gibt es Überraschungen?
      * Gibt es ungewöhnliche Muster/ Erklärungen für das Verhalten?
    - Ungewöhnliche Werte:
      * Ausreißer können durch coord\_cartesian sichtbar gemacht werden

**Fehlende Werte**

* Optionen:
  + Entfernen der Observationen
  + Anpassung der Variable mit mutate und bedingten Funktionen: Werte ersetzen

**Kovarianz**

* Beschreibt das Verhalten zwischen mind. zwei Variablen 🡪 Beziehung der Ausprägung zwischen den Variablen
  + Nominal-Metrisch
    - Bei Gruppierung und der Beziehung zwischen nominalen und metrischen Daten kann die Dichteverteilung in der Gruppe betrachtet werden
      * Histogram mit freqpool
      * Boxplot
    - Kann mit reorder() in aes nach Gruppe und Eigenschaften sortiert werden
  + Nominal-Nominal
    - Vergleich der Häufigkeit der Variablen-Kombinationen
      * Geom\_count
      * Geom\_tile 🡪 Kacheldiagram
        + Libraries: heatmaply oder d3heatmap
  + Metrisch-Metrisch
    - Scatterplot
      * Ergänzt mit bin2d kann die Dichte mit einbezogen werden altern. geom\_hex
    - Boxplots

**Pattern und Modelle:**

* Werden erarbeitet sofern eine Beziehung/Kovarianz zwischen Variablen festgestellt wurde
* Modelle extrahieren die Muster aus den Daten und bilden sie ab
  + Dienen der Vorhersage oder Klassifikation
  + Können über modlr erzeugt und analysiert werden

Mehr über ggplot: <https://amzn.com/331924275X> oder Grafiken: <https://amzn.com/1449316956>

**Kapitel 28: Graphics for communication**

https://r4ds.had.co.nz/graphics-for-communication.html