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>

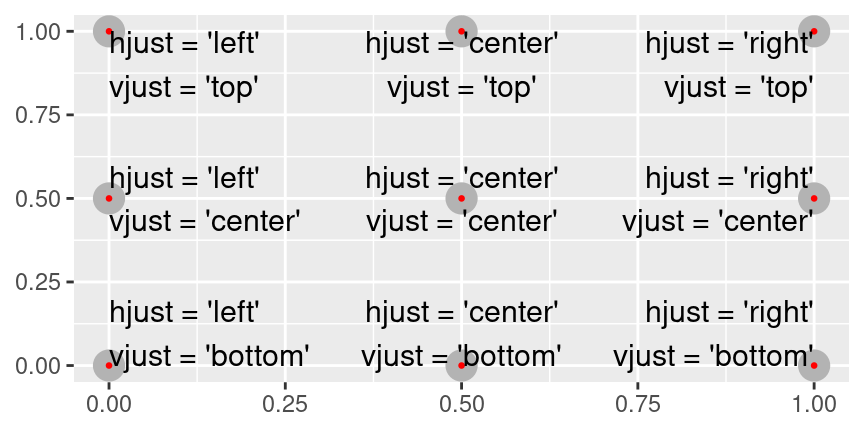
**Kapitel 28: Graphics for communication**

**Label**

* Titel:
  + Bsp. labs(title = "Fuel efficiency generally decreases with engine size")
  + Tipp: Resultat im Titel beschreiben, statt um was für eine Grafik es sich handelt
* Untertitel
* Randbemerkungen/ Caption
* X-/Y-Achse
* Legende bei Gruppierung über den zu gruppierenden Parameter z.B. Color, Shape o.ä.
* Mit quote()- können auch mathematischen Funktionen als Beschriftung ergänzt werden
  + ?quote()
  + ?plotmath

Annotations

* Benennung einzelner Datenpunkte
  + Geom\_text 🡪 geeignet zum Einfügen eines Labels oder Textblocks
  + Geom\_label 🡪 Besser als geom text, da keine Überlappung mit Bildpunkten, aber mit Labels
  + ggrepel::geom\_label\_repel 🡪 Überlappende Labels werden korrigiert angezeigt
* Hinzufügen einer seitlichen Anmerkung:
  + Summarise() + geom\_text(aes(label = label), data = label, vjust = "top", hjust = "left", nudge\_x=-1.15) siehe script LE4\_28



* + Aufteilen des Textes z.B. mit stringr::str\_wrap(width = 40) %>% writeLines()
  + Alternativ auch mit annotate und den Koordinaten möglich
* Weitere Möglichkeiten:
  + Geom\_hline/-vline: Liniengröße
  + Geom\_rect: Rechtecke um Bildpunkte zum Hervorheben
  + Geom\_segment() + arrow-Argument 🡪 Zeichnet Pfeil zu bestimmten Datenpunkt
* Annotation für einzelne Facets:
  + <https://stackoverflow.com/questions/11889625/annotating-text-on-individual-facet-in-ggplot2>

**Skalen/Scales**

* Verschiedene Ausprägungen mit scale\_ Bsp. **scale\_**y\_discrete
* Möglichkeiten:
  + Kontrollieren das Mapping der Daten im Graphen
  + Achsenbeschriftung der Einheiten unterdrücken mit labels=Null
  + Breaks, ermöglichen es den Abstand auf der Achse nach Belieben zu verändern
  + Legenden Layout über theme(legend.position=“alignment“) zu verändern oder auszublenden
  + Austauschen von Skalen entweder die Achsen oder die Färbungs/Shapeskala
    - Scale\_colour\_brewer(palette = „Set1“)
    - scale\_colour\_manual(values = c()) # Für eigene Farbauswahl, statt colour auch \_fill\_
    - scale\_colour\_viridis() # geeignet für Farbverläufe

**Zooming:**

* **Anpassung der Graphenbeschränkungen über** 
  + Eingabedaten
  + Skalenlimits
  + Limitierung des Koordinatensystems mit xlim/ylim

**Themes:**

* Nicht Datenelemente können mit Themes angepasst werden 🡪 Außendarstellung der Graphen 🡪 library(ggthemes)
* Auch eigene Themes können erstellt werden

**Speichern von Plots**

* Ggsave oder knitr
* In R-Markdown speichern:
  + Größenoptionen:
    - fig.width und/oder fig.height,
    - fig.asp,
    - fig.align, Einbettung im Text
    - out.width
    - out.height

Kapitel 6 Statistik mit R

**Repetetorium:**

* Begriffsdefinition Ökonometrie:
  1. Einsatz von Regressionsmethoden im ökonomischen Bereich
* Unterschied der Regression zur Korrelation
  + Korrelation bestimmt nur die stärke eines Zusammenhangs, aber dient nicht der Bildung eines Schätz-/Prognosewerts
  + Regression fragt nach einer guten Kombination zwischen verschiedenen Parametern zur Prognose einer Variable
* Interpretationseinschränkungen der Regression:
  + Interpretation geht nicht über die Modellgrenzen hinaus 🡪 anderes Wachstumsverhalten möglich
  + Regression beschreibt nur den Zusammenhang, aber nicht die Kausalität
    - Die abhängige Variable könnte umgekehrt auch die unabhängige Variable vorhersagen

**Annahmen der linearen Regression:**

1. Der wahre Zusammenhang zwischen erklärender/unabhängiger Variable und abhängiger Variable ist linear beschreibbar:
   1. Mit Steigungsparamtern auf Basis der Ceteris-paribas-Interpretation (wenn alle anderen Parameter konstant sind)
   2. Mit einem Startparamter/Achsenabschnitt, wenn alle unabhängigen Variablen den Wert 0 haben
   3. Störgröße u, die als Residuum bezeichnet wird und die Genauigkeit der Regressionsfunktion beeinflusst
2. Der Erwartungswert der Störgröße ist 0
   1. Im Mittel ist die Störgröße bei gleicher/wiederholender Stichprobe 0
3. Homoskedastizität der Störvariable
   1. Varianz der Störgröße ist an allen Messpunkten gleich
4. Keine Autokorellation der Störgröße
   1. Störgrößen einzelner Beobachtungen sind unabhängig voneinander und beeinflussen sich nicht
5. Die Störgröße ist aufgrund ihrer Unabhängigkeit normalverteilt
6. Unabhängige Variablen sind exogen und nicht stochastisch
   1. Sie besitzen feste Werte in der Regressionsmodellierung
   2. Nur durch die Zufalls- und Störgröße u wird Y zur Zufallsvariable
7. Kein perfekter Zusammenhang zwischen unabhängigen Variablen
   1. Es existiert keine perfekte Multikollinearität, ansonsten liegt eine Redundanz in der Bestimmung der abhängigen Variable vor
8. unabhängige Variablen besitzen Variation
   1. Ohne Variation keine Modellentwicklung
   2. Nur durch Veränderung zwischen x1 und x2, kann eine Veränderung bei y prognostiziert werden