Grundlagen der Datenvisualisierung

# LE1: Grundlagen der Visualisierung und Diagrammtypen

## Einleitung:

Ein Data Scientist hat eine Vielzahl von Aufgaben, die ein breites Spektrum an Fähigkeiten erfordern. Ein erheblicher Teil seiner Arbeit besteht aus der Aufbereitung und Analyse von diversen Datenquellen. Dabei ist jedoch nicht nur die Analyse selbst von Bedeutung, sondern auch die fachgerechte Interpretation und ansprechende Visualisierung der Ergebnisse. Die Auswahl der geeigneten Visualisierungsformen ist hierbei entscheidend, um die zugrundeliegenden Informationen klar und verständlich zu vermitteln. Nicht jedes Diagramm eignet sich gleichermassen zur Darstellung unterschiedlicher Informationsarten. Diese Lerneinheit beschäftigt sich mit den verschiedenen Diagrammtypen und den Kontexten, in denen sie am effektivsten eingesetzt werden können.

## Datenset:

Um die unterschiedlichen Diagrammtypen zu illustrieren, wird das «taxi» Dataset aus der Python Library «Seaborn» verwendet. Dieses Set beinhaltet Daten zu Taxifahrten, einschließlich Start- und Endzeitpunkten sowie den jeweiligen Ortschaften. Zusätzliche Informationen zur Fahrt, wie Preis, Zahlungsmethode oder Trinkgeld, sind ebenfalls verfügbar.

## Ein Bild, das Text, Screenshot, Reihe, Diagramm enthält. Automatisch generierte BeschreibungBeispiele:

Im Plot links wurde versucht der Fahrpreis in Abhängigkeit der gefahrenen Strecke abzubilden. In diesem Fall ist ein Scatterplot gut geeignet diese Beziehung darzustellen, da beide Attribute kontinuierlich sind.  
Dabei entspricht jeder Punkt einer Observation im Datenset.   
Um Regionen, wo viele Punkte aufeinanderliegend sind, sichtbar zu machen, wurde die Transparenz der Punkte erhöht.

Aus dem Diagramm lassen sich nun verschiedene Beobachtungen entnehmen.

Zum Beispiel lässt sich erkennen, dass zwischen den beiden Variablen ein positiver möglicherweise linearer Zusammenhang besteht (orange Linie).

Ein Bild, das Text, Screenshot, Reihe, Diagramm enthält.

Automatisch generierte BeschreibungDes Weiteren lässt sich auch erkennen, dass einige Observationen auf ca. 50CHF begrenze sind (Strecke zwischen 15 und 23km). Zusätzlich lässt sich auch erkennen, dass einige Taxis einen grossen Betrag einfordern für eine Strecke von 0km.

Des Weiteren wird hier noch das Preis/Leistung Verhältnis untersucht.

Dazu wurde beobachtet, ob sich die Strecke pro CHF über die Gesamtstrecke der Taxifahrt verändert.

Auch dies lässt sich mit einem Scatterplot gut darstellen. Erkennbar ist ein nichtlinearer positiver Trend abgesehen von ein paar Ausreissern.

Ein Bild, das Text, Screenshot, Diagramm, Reihe enthält.

Automatisch generierte BeschreibungNun möchten wir Rückschlüsse auf die Verteilung der gefahrenen Strecke ziehen.

Mean, Median und Standardabweichung sind dabei nicht unbedingt ausreichend, um sich ein Bild der Verteilung machen zu können.

Dazu gibt es Visualisierungsarten wie beispielsweise das nebenstehende Histogramm.

Dieses Histogramm erlaubt uns ein Überblick zu verschaffen.

Beispielsweise kann man sehen, dass die meisten Taxifahrten um die 2-3 km lang wahren.   
Es gab allerdings auch ein paar wenige Fahrten, welche zwischen 20 und 40 km lang waren.

Wir konnten uns nun ein grobes Bild der Verteilung machen.  
Diese kurzen bzw. langen Fahrten könnten möglicherweise davon abhängen, wo genau ein Kunde eingestiegen ist.

Ein Bild, das Text, Diagramm, Screenshot, Rechteck enthält.

Automatisch generierte BeschreibungDeshalb wurden die Taxifahrten nach dem Startbezirk (Bezirk, wo die Fahrt begonnen hat) gruppiert und anschliessend Boxplots gezeichnet.

Boxplots können wie Histogramme, Auskunft über die Verteilung geben.  
Im Gegensatz zu den Histogrammen tun diese dies allerdings mittels der vier Quantile.

Im nebenstehenden Plot sieht man also nun vier Boxplots gruppiert nach dem Startbezirk.

In Manhattan sowie Brooklyn als Startbezirk werden meist kürzere Strecken gefahren (wenn man die Ausreisser ignoriert).  
Fahrten, welche in Queens beginnen, sind in der Regel am längsten.  
Alle Regionen beinhalten Fahrten mit einer Strecke von 0 km.

Die vielen Ausreisser der Boxplots lassen darauf schliessen, dass die Verteilungen wie im Histogramm ersichtlich, eher langschwänzig sind.

Das Histogramm sowie die Boxplots sind Diagrammarten, welche für Laien eher schwierig zu interpretieren sind, da sie Vorwissen über Verteilungen oder die Erschaffungsweise der Diagramme erfordern. Daher sind solche eher für ein Expertenpublikum geeignet.  
Auch als Experte muss man aufpassen, da beispielsweise der Boxplot keine eindeutige Aussage über den Skewness / Verteilungsart machen kann.

Bei einem Laienpublikum sollte man eher auf Scatterplots, Barcharts und weitere Diagrammarten zurückgreifen, welche ohne Vorwissen intuitiv verständlich sind.

# LE2: Visuelle Wahrnehmung

## Einleitung:

Dieses Thema beleuchtet die Art und Weise, wie Menschen visuelle Informationen verarbeiten und interpretieren, und erforscht, wie diese zur Vermittlung von Informationen effektiv eingesetzt werden können. Im Zentrum der Betrachtung stehen dabei die Gesetze der Gestaltwahrnehmung, welche beschreiben, wie Menschen unterschiedliche Elemente als organisierte Strukturen erfassen. Die Kenntnis dieser Gesetze ist entscheidend, um Diagramme zu erstellen, die leicht verständlich und auf den menschlichen Verstand abgestimmt sind.

Ein weiterer bedeutender Aspekt, der berücksichtigt werden muss, ist der «Cognitive Load», also wie viele Informationen in einem Diagramm dargestellt werden und wie diese verschachtelt sind. Um den genannten Aspekten gerecht zu werden, kann man auf die sieben Variablen von Bertin zurückgreifen: Position, Größe, Form, Farbe, Helligkeit, Orientierung und Textur.   
Im weiteren Verlauf werden einige dieser Variablen näher analysiert und Beispiele für ihre Anwendung dargestellt.

## Datenset:

Um einige der Variablen von Bertin zu analysieren, wird das Iris Datenset verwendet.  
Die in Iris natürlich vorkommenden Cluster eignen sich perfekt, um Gruppierungen verschieden darzustellen.  
Das Iris Datenset beinhaltet vier Features (sepal\_width/length, petal\_width/length).  
Petal und Sepal referenzieren dabei auf das Blütenblatt respektive Kelchblatt.

## Beispiele:

### Ein Bild, das Text, Screenshot, Diagramm, Reihe enthält. Automatisch generierte BeschreibungPosition und Farbe:

Im nebenstehenden Scatterplot wurden die Dimensionen der Blütenblätter verschiedener Blumen geplottet. Um die Spezies unterschieden zu können, wurden die einzelnen Blumen nach deren Zugehörigkeit eingefärbt.   
Wir können ausgehend der Dimensionen des Blütenblattes nun erkennen, dass sich «versicolor» viel näher an «virginica» befindet als «setosa». Auch lassen sich so die einzelnen Spezies als zusammengehörige Gruppierungen erkennen.

Ein Bild, das Text, Screenshot, Diagramm, Reihe enthält.

Automatisch generierte Beschreibung

Hier wurde nochmals derselbe Plot erzeugt, allerdings ohne die einzelnen Spezies einzufärben.  
Nun lassen sich nur noch zwei Gruppierungen erkennen. «versincolor» sowie «virginica» lassen sich voneinander nicht mehr unterscheiden, da diese laufend ineinander verschmelzen.   
In diesem Plot ist es also unerlässlich, dass eine weitere Variable eingesetzt wird.

Berins Variable der Position kann mit dem Gesetz der Nähe korrelieren, da Elemente, welche nahe beieinander liegen als zusammengehörig wahrgenommen werden können.  
Um dieses Gesetz zu bestärken, kann eine weitere Variable wie Farbe/Helligkeit oder auch die Form der Datenpunkte verwendet werden.

Möchte man verschiedene Farben einsetzen, so ist es wichtig Farbenblinde zu berücksichtigen.  
Im Zweifelsfall setzt man lieber auf die Helligkeit (z.B. verschiedene Graustufen) oder die Form.

### Ein Bild, das Text, Screenshot, Diagramm, Schrift enthält. Automatisch generierte BeschreibungGrösse:

Wenn man in einer Visualisierung den Fokus auf eine bestimmte Sache legen möchte, so kann dies mit der Grösse eines Objektes erreicht werden.

Objekte welche grösster sind oder sich auch im Vordergrund befinden werden meist als wichtig empfunden und lenken die Aufmerksamkeit auf sich.

Im nebenstehenden Plot wurden die Datenpunkte von «virginica» vergrösstert dargestellt.

### Ähnlichkeit:

Ein Bild, das Text, Screenshot, Diagramm, Reihe enthält.

Automatisch generierte BeschreibungWie bereits mit den Farben, kann auch mit den Formen der Elemente gespielt werden.

Hier wurde für jede Spezies eine andere Form der Datenpunkte gewählt. Auch so kann erreicht werden, dass der Betrachter die einzelnen Klassen unterscheiden kann.

Das menschliche Gehirn nimmt ähnliche Formen als zusammengehörig wahr.

### Cognitive load:

Ein Bild, das Text, Screenshot, Diagramm, Schrift enthält.

Automatisch generierte BeschreibungDer «Cognitive Load» beschreibt die Menge an Informationen, die unser Arbeitsgedächtnis verarbeiten kann.

Visualisierungen mit einer Menge verschachtelter Informationen sind ein grösserer «Cognitive Load» als solche mit weniger Informationen und brauchen somit je nach Betrachter mehr Zeit, um verstanden zu werden (Beispiel für eine Visualisierung mit grösserem Cognitive Load siehe Abbildung links).

Deshalb ist es von grosser Wichtigkeit die Visualisierungen auf die Zielgruppe und Einbettungsbereich (Paper, Präsentation, Bericht usw.) anzupassen.