Grundlagen der Datenvisualisierung

# LE1: Grundlagen der Visualisierung und Diagrammtypen

## 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.

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 schon eher langschwänzig sind.

Das Histogramm sowie die Boxplots sind Diagrammarten, welche für Laien eher schwierig zu interpretieren sind und daher eher für ein Expertenpublikum geeignet ist.  
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.