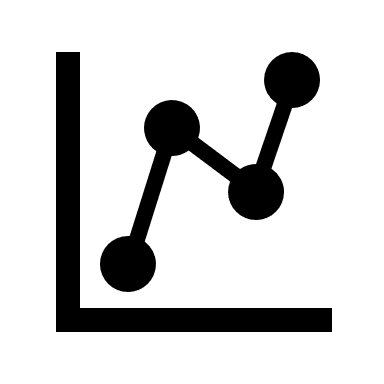
Grundlagen der Datenvisualisierung



Tobias Buess

Semester 5

# 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:

Scatterplots eignen sich gut um Beziehungen zwischen zwei kontinuierlichen Attributen herzustellen, da die Daten so als Punkte dargestellt werden können [1].

Im linken Plot habe ich versucht die Beziehung zwischen Strecke und Fahrpreis darzustellen. In diesem Fall ist ein Scatterplot gut geeignet diese Beziehung darzustellen, da beide Attribute kontinuierlich sind.  
  
Um Regionen, wo viele Punkte aufeinanderliegend sind, sichtbar zu machen, wurde die Transparenz der Punkte erhöht.

Abbildung : Strecke vs Fahrpreis

Es ist erkennbar, dass zwischen den beiden Variablen ein positiver möglicherweise linearer Zusammenhang besteht (orange Linie). Des 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.

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

Automatisch generierte BeschreibungDes 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.

Abbildung : Stecks vs Preis/Leistung

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 [2].

Dieses Histogramm erlaubt uns ein Überblick über die Verteilung zu verschaffen.

Abbildung : Gefahrene Strecke

Dabei teilt das Histogramm den Bereich, in welchem sich die Daten befinden in x gleich grosse Bereiche (Bins) auf, wobei anschliessend die Anzahl Datenpunkte gezählt wird, welche sich in einem Bin befinden. Jeder Bin wird als einzelne Bar angezeigt. In unserem Fall wurden die Counts auf der y-Achse als Dichte angezeigt. Histogramme eignen sich z.B. für kontinuierliche Attribute wie es bei uns der Fall ist.

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 von kontinuierliche Daten geben.  
Im Gegensatz zu den Histogrammen tun diese dies allerdings mittels der vier Quantile [3].  
Auf dem Boxplot ersichtlich sind Median, Q1-4 und diverse Ausreisser.

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

Abbildung : Gefahrene Strecke in abh. vom Startort

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 [4] 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 [5].

## 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:

Hier wird die Variablen Position und Gestalt behandelt.   
Diese erlauben es uns Elemente als Gruppen wahrzunehmen, wie im Beispiel links erkennbar ist.   
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 [6].

Abbildung : Blattbreite vs Länge

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.

Abbildung : Blattbreite vs Länge (einfarbig)

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 Variable «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 [5].

Abbildung : Blattbreite vs Länge (virginica big)

### 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 [5].

Abbildung : Blattbreite vs Länge (Form Art)

### 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» [5] 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.

Abbildung : Blattbreite vs Länge (KDE)

# LE3: Designprinzipien vs. Daten

## Einleitung:

Daten werden meist nicht von den Personen erhoben, welche sie schlussendlich auch Analysieren und Visualisieren. Im heutigen Zeitalter von «Big Data» werden auch preisgünstig alle möglichen Daten erhoben und abgespeichert in der Hoffnung, dass man aus diesen, zu einem späteren Zeitpunkt, irgendwelche Erkenntnisse gewinnen kann. Dies führt oft dazu, dass gesammelte Daten unstrukturiert abgespeichert sind und diese erstmals gefiltert und vorverarbeitet damit sie verwendet werden können[7].

Des Weiteren kommt es auch öfters vor, dass während der Gewinnung von Daten etwas schiefläuft und Observationen unerlaubte Werte annehmen, mehrfach aufgezeichnet werden oder sogar ganz ausbleiben. Deshalb ist es sehr wichtig, dass immer eine Bereinigung der Daten und ein Sanity Check (machen die Werte der Observationen Sinn) durchgeführt wird.

Ein weiterer wichtiger Punkt, welcher vor allem die Visualisierung betrifft, ist die Art, in welcher die Daten vorliegen. Denn nicht jede Art von Datenpunkten können auf dieselbe Art und Weise dargestellt werden.  
Daten lassen sich dabei in die Kategorien stetig, diskret, ordinal und nominal einordnen [8].  
Beispiele für die Datentypen und deren geeignete Visualisierungen werden nachfolgend aufgezeigt.

## Daten:

In diesem LE wird folgendes Datenset verwendet:  
<https://www.bfs.admin.ch/bfs/de/home/statistiken/energie.assetdetail.28065403.html>

In diesem befinden sich die Daten über den Endverbrauch der erneuerbaren Energien der Schweiz vom Jahre 1990 bis 2022. Hier war kein exzessives Data-Cleaning nötig, da diese bereits als Tabelle vorliegen (Excel Tabelle). Allerdings mussten irrelevante Zellen entfernt werden.

## Beispiele:

### Darstellung stetiger Werte:

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

Automatisch generierte BeschreibungStetige Daten oder auch kontinuierliche Daten gehören zu der Gruppe der Quantitativen Daten.  
Stetige Daten können auf einem bestimmten Intervall theoretisch unendlich viele Werte annehmen.  
Ein Beispiel dafür wäre die Temperatur, denn diese kann zwischen dem absoluten Nullpunkt bis zur Planck Temperatur jeden beliebigen Wert annehmen.

In diesem Beispiel wurde der Verlauf der verbrauchten, durch die Sonne generierte Energie, visualisiert.

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

Automatisch generierte BeschreibungDie Energie ist in der Einheit TJ (Terra Joule) angegeben. Terra Joule ist eine stetige Mengenangabe und kann in unserem Fall mittels eines Line Plots visualisiert werden.

Abbildung : Verlauf Energieverbrauch

Nun möchten wir die Zunahme der Sonnenenergie analysieren. Dazu wurde lediglich die die Differenz zum Vorjahr berechnet und visualisiert.

Diesmal wurde ein Barplot verwendet, da dies sich hier anbietet, da x diskret und y stetig ist. Aus diesem Plot kann man nun herauslese, dass der Anstieg im Jahre 2010 am grössten war.

Abbildung : Verlauf Differenz Energieverbrauch

### Darstellung diskreter Werte:

Diskrete Daten gehören wie die stetigen auch zu der Gruppe der quantitativen Daten.  
Sowohl stetige als auch diskrete Daten (quantitative Daten) sind somit Zahlen, bei welchen ein Abstand zwischen zwei Punkten berechnet werden kann.

Im Gegensatz zu den stetigen Daten können diskrete Daten nicht unendlich viele Werte auf einem Intervall annehmen.   
Dazu gehören beispielsweise diverse Aufzählungen oder Nummerierungen.

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

Automatisch generierte BeschreibungBei dem Kapitel mit den stetigen Werten wurde bereits den Umgang mit diskreten Datentypen in der Visualisierung angedeutet.

Denn tatsächlich ist die x-Achse mit dem Jahr bereits diskret.

In diesem Plot wurden allerdings mehrere Energiearten dargestellt, weshalb hier nicht mehr jedes Jahr einzeln visualisiert werden kann.

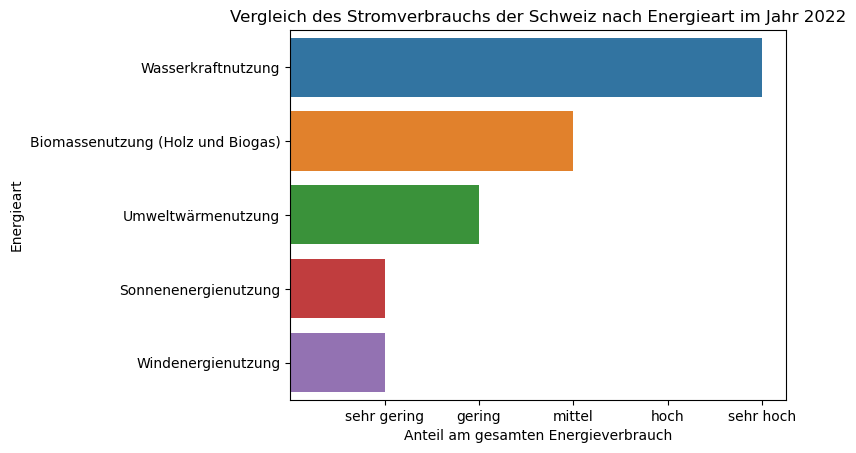
Theoretisch hätte man auch ein Lineplot mit jeweils einer Linie pro Art verwenden können.

Abbildung : Energieverbrauch in Intervallen

Hier wurde nun ein Binning auf das Jahr angewendet und dabei die Energiemengen der Innerhalb eines Bins aufaddiert. Somit erhalten wir die verbrauchte Energiemenge innerhalb eines bestimmten Intervalls. Die Bins sind alle in etwa gleich gross.

Somit konnten wir nun Informationen komprimiert darstellen. Allerdings ist dabei wichtig zu beachten, dass die Bin-grösse so gewählt wird, dass möglicherweise wichtige Informationen nicht verloren geht. Dies ist allerdings Use Case abhängig.

## Darstellung ordinaler / nominaler Daten:

Ordinale bzw. nominale Daten gehören zu der Gruppe der qualitativen Daten.  
Diese sind nun nicht mehr zwingend Zahlen, sondern können auch Kategorien sein, bei welchen kein Abstand quantifizierbar ist.

In diesem Beispiel wurde die Menge an verbrauchtem Strom des Jahres 2022 einiger ausgewählten Energiearten visualisiert. Dabei wurden die tatsächlichen Energiemengen durch eine qualitative Menge ersetzt.

Die Zuordnung wurde durch ein Binning gelöst, wobei so ein ordinaler Datentyp entstand.   
Ordinal bedeutet, dass es sich um ein Kategorieller Datentyp handelt, dieser allerdings logisch geordnet werden kann.   
Ob diese Abstufung Sinn macht, kommt immer auf den Use Case an. Beispielsweise kann so eine Variable für ein Laienpublikum «greifbarer» gemacht werden.

Abbildung : Stromverbrauch nach Energieart

Zu den nominalen Daten gehören Kategorien, welche nicht in einer logischen Reihenfolge geordnet werden können.

# LE4: Grammar of Graphics

## Einleitung:

Die 'Grammar of Graphics' (GoG) stellt eine Herangehensweise an die Datenvisualisierung dar.   
Anstatt traditionelle, festgelegte Diagrammtypen in den Vordergrund zu stellen, bietet GoG ein modulares Framework, welches auf einem layered approach basiert.   
Diese Methodik erlaubt es, diverse grafische Elemente in strukturierten Schichten zu kombinieren, um Daten in einer strukturierten Weise auch auf mehreren Dimensionen darzustellen [9].  
Folgend werden einige der sieben Layer aufgezeigt und erläutert was hinter ihnen steckt.

## Daten:

Hier wird das «diamonds» Datenset vom Python Package «plotnine» verwenden.  
Dieses Datenset umfasst ca. 54'000 Datenpunkte und beinhaltet verschiedene Attribute von verschiedenen Diamanten.

## Beispiele:

### Layers: Data & Aesthetics & Geometric objects:

Dies sind die Grundbausteine des Grammar of Graphics und werden immer benötigt, um ein Plot zu erstellen [10].

Dabei brauchen wir als erstes Daten mit mindestens einer Dimension, welche wir darstellen möchten.

Anschliessend müssen wir festlegen, wie genau die verschiedenen Dimensionen visualisiert werden sollen. Dazu muss ein Mapping von den Attributen der Daten auf verschiedenen Aestetics erfolgen.  
Dazu gehören beispielsweise die Position (x, y-Achse), Farbe, Grösse usw.

Des Weiteren brauchen wir auch noch ein geometrisches Objekt, welches schlussendlich die Datenpunkte repräsentiert.  
Dies können Punkte (Scatterplot), Bars (Barchart), Linie (Lineplot) usw. sein.

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

Automatisch generierte Beschreibung In diesem Plot wurde zunächst nur der Preis in Anhängigkeit des Gewichts geplottet.

Bezüglich Layers des GoG können wir hier einerseits die Daten erkennen, sowohl auch die Aestetics mit dem Mapping der Daten auf die x und y-achse.

Zu guter Letzt sehen wir logischerweise auch die Punkte als geometrisches Objekt.  
Dabei wurde die Transparenz der Punkte erhöht, damit dichtere Regionen im Plot gut erkennbar sind.

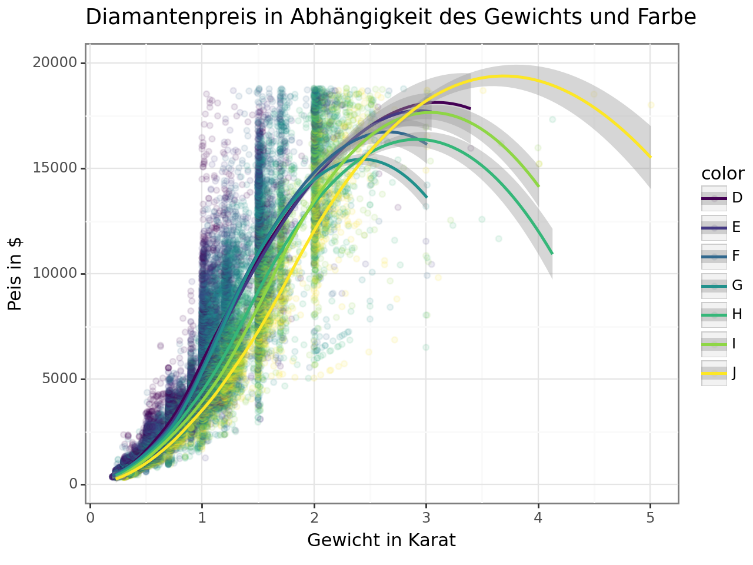
Abbildung : Diamantenpreis nach Gewicht

Anhand dieses Plots kann ein erster Eindruck gewonnen werden.   
Dabei ist erkennbar, dass schwerere Diamanten logischerweise teurer sind.  
Allerdings ist auch eine trichterförmige Form der Datenpunkte zu erkennen.  
Schwerere Diamanten streuen also viel stärker im Preis als leichtere.

Des Weiteren ist dem Plot zu entnehmen, dass Diamanten unter einem Karat gleichmässig verteilt sind, verglichen zu den einzelnen Ansammlungen um 1, 1,5 und 2 Karat.

### Layers: Aesthetics & Geometric objects & Statistics:

Um einen Überblick über die Gruppierungen zu erhalten und möglicherweise etwas über die Preise zu lernen, können wir weitere Aesthetics wie zum Beispiel die Farbe verwenden.

Wie man allerdings sieht, kann man einzelne Klassen nicht wirklich erkennen, da einfach zu viele Datenpunkte vorhanden sind.

Um die Unterschiede trotzdem etwas besser zu erkennen, wurde eine Statistik hinzugefügt.

Diese Statistik zeigt jeweils eine Locally Weighted Regression mit Konfidenzintervall nach Diamantfarbe.

Die Regressionen sind unterschiedlich, weshalb somit die Diamantfarbe einen möglichen Einfluss auf den Preis hat.

Abbildung : Diamantenpreis nach Gewicht und Farbe

Um die Klassen etwas besser unterscheiden zu können, gibt es ein weiterer Layer im GoG, welcher wir nutzen können:

### Layer: Facets:

Bei diesem Layer handelt es sich um Facets.  
Beim Faceting handelt es sich um ein Aufsplitten des Datensets, wobei diese Splits in einem separaten Plot visualisiert werden können [9].  
Diese «subplots» teilen sich üblicherweise die gleiche Achse(n) um die Vergleichbarkeit der einzelnen Plots beizubehalten.

Ein Bild, das Diagramm, Text enthält.

Automatisch generierte Beschreibung

Abbildung 16: Diamantenpreis nach Gewicht und Farbe (Facets)

In diesem Plot wurde ein facet wrap verwendet, wobei jedes Subplot einer Farbe entspricht.  
Auf den jewiligen Subplots sind nun wirklich nur die Datenpunkte vorhanden, welche zu der entsprechenden Klasse gehört. Dadurch wird der Plot leserlicher und kann besser interpretiert werden. In ggplot gibt es auch ein facet grid Objekt, mit welchem eine weitere Dimension dargestellt werden kann.

# LE5: Evaluation

## Einleitung:

Visualisierungen kann man für sich selbst zum Verstehen von Daten oder auch für andere Personen erstellen.   
Bei letzterem ist üblicherweise das Ziel, dem Betrachter bestimmte Informationen zu vermitteln.  
Doch wie kann man sicher sein, dass die Visualisierung genau so verstanden wird, wie es ursprünglicherweise geplant war? Oder ist die Visualisierung gar zu kompliziert?

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

Automatisch generierte BeschreibungUm solche Fragen zu beantworten, gibt es Usability-Tests.   
Die Ziele von Usability-Tests sind vielfältig und variieren zwischen Anwendungsbereichen, beinhalten allerdings in den meisten Fällen folgende Punkte:

[11]

Usability-Tests können sowohl qualitativ als auch quantitativ sein.

Bei quantitativen Tests werden Metriken erhoben wie «task success» und «time on task» um die Benutzererfahrung zu beschreiben.

Qualitative Tests konzentrieren sich auf das Sammeln von Einsichten wie der Benutzer ein Produkt nutzt und um Probleme zu identifizieren [11].  
Dabei werden einem Benutzer der Zielgruppe verschiedene Aufgaben vorgelegt, welcher er lösen muss. Eine Aufsichtsperson leitet dabei den Test und beobachtet den Benutzer und gibt Instruktionen. Dabei ist es wichtig, dass die Aufsichtsperson den Benutzer nicht beeinflusst.

Um Visualisierungen zu testen, kann ein Usability-Test nun wie folgt aussehen:

1. Definieren, welche Fragestellung mit dem Usability-Test beantwortet werden soll.
2. Entwickeln von Fragen oder Tasks, welche der Benutzer beantworten / lösen soll.  
   Dabei ist es wichtig, dass die Tasks so gewählt werden, dass der Benutzer so wenig beeinflusst wird wie möglich.
3. Test mit Testperson durchführen und Verhalten sowie Antworten analysieren
4. Rückschlüsse ziehen, Grafik verbessern und Test mit neuer Gruppe von Testpersonen durchführen.

Bei dieser Gruppe von Testpersonen handelt es sich um ca. 5 Personen aus einer geeigneten Zielgruppe.  
Wenn eine Gruppe zu wenig Personen enthält, kann es sein, dass wichtige Informationen und Feedback nicht erfasst wird. Wenn allerdings die Gruppe zu gross ist, erhält man mit der Zeit nur noch wenige neue Informationen, da sich das Feedback mit der Zeit immer mehr mit dem bereits erhaltenen überschneidet [12].

Durch das Feedback der Testpersonen, kann anschliessend eine Visualisierung entsprechend angepasst werden.

Dieses Testing ist ein iterativer Prozess, wobei immer eine Gruppe getestet, anschliessend wie Visualisierungen verbessert und anschliessend der Test mit einer neuen Gruppe ausgeführt wird.

## Beispiel:

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

Automatisch generierte BeschreibungEin Bild, das Text, Screenshot, Diagramm, Reihe enthält.

Automatisch generierte BeschreibungEs gibt viele verschiedene Arten von Usability-test. Einer davon ist der A/B Test.  
Dieser eignet sich für Tasks, bei welchen zwischen zwei Versionen von beispielsweise einer Website entschieden werden muss. Folgend möchte ich also mittels A/B Test, testen, welche der zwei Grafiken aus LE2 besser geeignet sind, um die enthaltenen Informationen darzustellen.

Beide Darstellungen enthalten dieselben Informationen. Während in der linken Grafik die Datenpunkte farblich nach Spezies differenziert sind, weisen sie in der rechten Grafik unterschiedliche Formen je nach Spezies auf.   
Ich möchte herausfinden, ob die rechte Darstellung ebenso intuitiv verstanden wird wie die linke. Falls nicht, möchte ich Wege finden, diese zu optimieren.   
Dabei verfolge ich das Ziel, komplett auf Farben zu verzichten, sodass auch farbenblinde Personen keinerlei Nachteile erleben.

### Tasks definieren:

* Rechte Grafik vorlegen und Fragen, welche Informationen daraus entnommen werden können.
* Linke Grafik hinzufügen und Fragen welche Grafik als angenehmer empfunden wird.
* Fragen, was bei der Schwierigen zu interpretierenden Grafik anders sein müsste, damit diese einfacher zu interpretieren ist.

### Gruppe von Testpersonen:

Ich habe eine Testgruppe aus fünf Personen aus meinem direkten Umfeld zusammengestellt.   
Keine der Testpersonen hat umfangreiche Erfahrung im Bereich Datenvisualisierung, was das Ergebnis möglicherweise beeinflussen könnte.

### Ergebnis / Verbesserungsmöglichkeiten:

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

Automatisch generierte BeschreibungDie Analyse zeigt, dass eine farbige Visualisierung als angenehmer wahrgenommen wurde.   
Da in der Testgruppe keine Farbenblinden waren, könnte dieses Ergebnis verzerrt sein. Die Legende war für die Tester verwirrend, da sie die speziellen Blumennamen nicht kannten.

Meine Verbesserungsvorschläge sind in der Grafik links ersichtlich. Ich behalte verschiedene Formen für die Gruppen bei, um Farbenblinde zu berücksichtigen.  
Jedoch färbe ich zusätzlich die Datenpunkte nach Gruppe, um den nicht Farbenblinden Personen eine vertraute Darstellung zu bieten.   
Die Legende bleibt unverändert, da die deutschen Namen ähnlich sind wie die bestehenden.   
Dabei klärt der Titel nun, dass es sich um Schwertlilien handelt, was die Legende verständlicher macht.

Abbildung : Verbesserung plot LE2

**Literaturverzeichnis:**

[1] J. Castañón, „10 Visualizations Every Data Scientist Should Know“, Medium. Zugegriffen: 5. Oktober 2023. [Online]. Verfügbar unter: https://towardsdatascience.com/10-viz-every-ds-should-know-4e4118f26fc3

[2] „How a Histogram Works to Display Data“, Investopedia. Zugegriffen: 5. Oktober 2023. [Online]. Verfügbar unter: https://www.investopedia.com/terms/h/histogram.asp

[3] „Box plot review (article) | Khan Academy“. Zugegriffen: 5. Oktober 2023. [Online]. Verfügbar unter: https://www.khanacademy.org/math/statistics-probability/summarizing-quantitative-data/box-whisker-plots/a/box-plot-review

[4] S. C. Government of Canada, „5.2 Bar chart“. Zugegriffen: 5. Oktober 2023. [Online]. Verfügbar unter: https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch9/bargraph-diagrammeabarres/5214818-eng.htm

[5] *Now You See It: Lessons from Research on Perception for Design of Data Visualizations*, (14. September 2017). Zugegriffen: 28. September 2023. [Online Video]. Verfügbar unter: https://www.youtube.com/watch?v=66eE4rc5xU0

[6] „Visual Variables“. Zugegriffen: 5. Oktober 2023. [Online]. Verfügbar unter: https://www.axismaps.com//guide/visual-variables

[7] I. C. Education, „Structured vs. Unstructured Data: What’s the Difference?“, IBM Blog. Zugegriffen: 7. Oktober 2023. [Online]. Verfügbar unter: https://www.ibm.com/blog/structured-vs-unstructured-data/

[8] „Guide to Data Types and How to Graph Them in Statistics - Statistics By Jim“. Zugegriffen: 7. Oktober 2023. [Online]. Verfügbar unter: https://statisticsbyjim.com/basics/data-types/

[9] D. (DJ) Sarkar, „A Comprehensive Guide to the Grammar of Graphics for Effective Visualization of Multi-dimensional…“, Medium. Zugegriffen: 7. Oktober 2023. [Online]. Verfügbar unter: https://towardsdatascience.com/a-comprehensive-guide-to-the-grammar-of-graphics-for-effective-visualization-of-multi-dimensional-1f92b4ed4149

[10] B. Soltoff, „The grammar of graphics“, Computing for Information Science. Zugegriffen: 7. Oktober 2023. [Online]. Verfügbar unter: https://info5940.infosci.cornell.edu/notes/dataviz/grammar-of-graphics/

[11] W. L. in R.-B. U. Experience, „Usability Testing 101“, Nielsen Norman Group. Zugegriffen: 8. Oktober 2023. [Online]. Verfügbar unter: https://www.nngroup.com/articles/usability-testing-101/

[12] *Usability Testing w. 5 Users: Design Process (video 1 of 3)*, (26. Oktober 2018). Zugegriffen: 8. Oktober 2023. [Online Video]. Verfügbar unter: https://www.youtube.com/watch?v=RhgUirqki50

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

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

Automatisch generierte BeschreibungEin Bild, das Text, Screenshot, Diagramm, Reihe enthält.

Automatisch generierte BeschreibungEin Bild, das Text, Diagramm, Screenshot, Rechteck enthält.

Automatisch generierte BeschreibungEin Bild, das Text, Screenshot, Diagramm, Reihe enthält.

Automatisch generierte Beschreibung

Abbildung 6: Blattbreite vs Länge (einfarbig)

Abbildung 5: Blattbreite vs Länge

Abbildung 4: Gefahrene Strecke in abh. vom Startort

Abbildung 3: Gefahrene Strecke

Abbildung 2: Stecks vs Preis/Leistung

Abbildung 1: Strecke vs Fahrpreis

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

Automatisch generierte BeschreibungEin Bild, das Text, Diagramm, Screenshot, Reihe enthält.

Automatisch generierte BeschreibungEin Bild, das Text, Screenshot, Diagramm, Schrift enthält.

Automatisch generierte BeschreibungEin Bild, das Text, Screenshot, Diagramm, Reihe enthält.

Automatisch generierte BeschreibungEin Bild, das Text, Screenshot, Diagramm, Schrift enthält.

Automatisch generierte Beschreibung

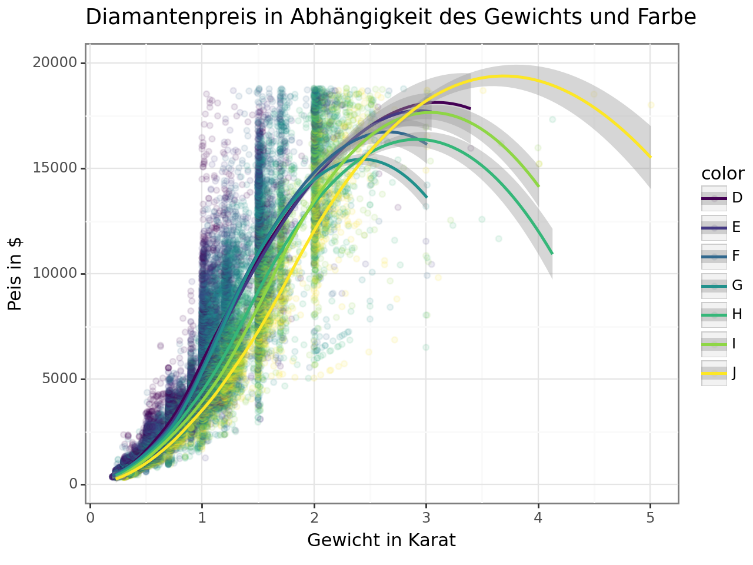
Abbildung 11: Verlauf Differenz Energieverbrauch

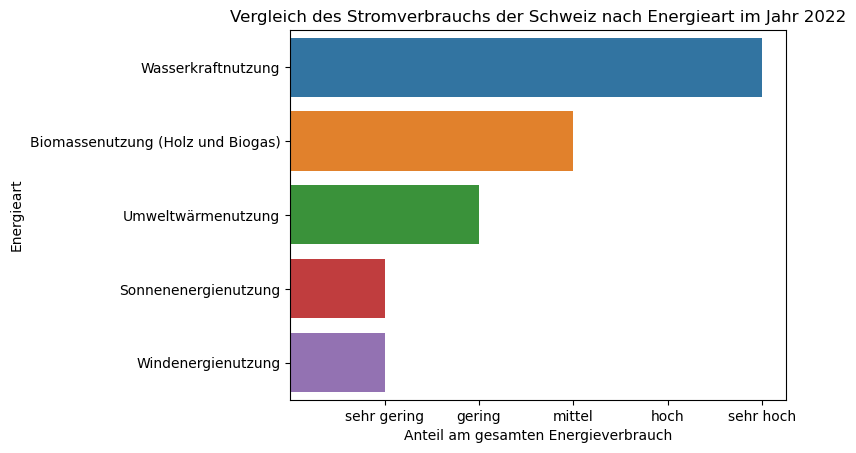
Abbildung 10: Verlauf Energieverbrauch

Abbildung 9: Blattbreite vs Länge (KDE)

Abbildung 8: Blattbreite vs Länge (Form Art)

Abbildung 7: Blattbreite vs Länge (virginica big)

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

Automatisch generierte BeschreibungEin Bild, das Text, Screenshot, Diagramm, Reihe enthält.

Automatisch generierte Beschreibung

Abbildung 15: Diamantenpreis nach Gewicht und Farbe

Abbildung 14: Diamantenpreis nach Gewicht

Abbildung 13: Stromverbrauch nach Energieart

Abbildung 12: Energieverbrauch in Intervallen

Ein Bild, das Diagramm, Text enthält.

Automatisch generierte Beschreibung

Abbildung 16: Diamantenpreis nach Gewicht und Farbe (Facets)

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

Automatisch generierte Beschreibung

Abbildung 17: Verbesserung plot LE2