Exploratory Data Analysis (EDA) – Erste Schritte

Datensatz: Entwicklung der Gehälter für Data Scientists 2020-2024

Der Datensatz enthält Informationen über Stellenbezeichnung, Beschäftigungsart, Erfahrungsstufe, Fachkenntnisstufe, Gehalt, Gehaltswährung, Unternehmensstandort, Gehalt in USD, Wohnort des Mitarbeiters, Unternehmensgröße und Jahr. Diese Informationen bieten einen umfassenden Überblick über berufsbezogene Details, Gehaltsniveaus, Unternehmensmerkmale und zeitliche Aspekte. Der Datensatz dient als Quelle für Personen, die eine Karriereberatung suchen, für Unternehmen, die ihre Vergütungsstrategien vergleichen möchten, und für Forscher, die die sich entwickelnde Dynamik des Arbeitsmarktes für Datenwissenschaftler untersuchen.

Datenquelle: https://ai-jobs.net/

Aufgaben:

1) Untersuchen Sie die Datenquelle! Was wird hier angeboten, woher kommen die Daten? Schätzen Sie die Verlässlichkeit der Daten ein!
2) Erstellen Sie ein Python-Programm, mit dem Sie den Datensatz für ihre weitere Analyse einlesen! Verschaffen Sie sich einen Überblick über den Datensatz. Wie viele Datensamples sind enthalten, wie viele Attribute gibt es? Gibt es fehlende Daten (z.B. „NaN“, „0“)? Zeigen Sie die ersten Zeilen des Datensatzes an! Erstellen Sie ein kurzes Info-Blatt zu den o.g. Metadaten!
3) Für einen ersten Überblick bieten sich statistische Kennzahlen an: Berechnen Sie das Gehalt in USD – Mittelwert und Median jeweils für den Gesamt-Datensatz und spezifisch für jede Kategorie in der Attribut-Gruppe „Experience.Level“!
4) Warum ist das Attribut „Salary.in.USD“ besser für die Analyse geeignet als das Attribut „Salary“?
5) Gibt es sogenannte „Ausreißer“ (Werte, die weit von der üblichen Werteverteilung abweichen) in der Verteilung der Gehälter der spezifischen Kategorien in der Attribut-Gruppe „Experience.Level“? Wie kann man so etwas grafisch darstellen?
6) Verändern sich die Gesamt-Gehälter im Verlauf der Jahre? Wie könnte man das überzeugend grafisch darstellen?
7) Welche Fragen könnte man anhand dieses Datensatzes weiter untersuchen?

- **Abgabe bis Mittwoch 15.01. 16:00 Uhr zwei Folien pro Gruppe (pdf!!), im Namen des Files bitte Gruppennummer bzw. Namen der Verfasser + Nummer der Übung verwenden (z.B. GruppeXX_uebung-a1.pdf)**
  - **in Verzeichnis Upload_DASC hochladen: https://nc.ufz.de/s/39zCZKFAyZKDQ2z** (Passwort: ?solution_DASC25)
- Folie 1: Infos zum Datensatz (Aufgaben 1- 3),
  Folie 2: Infos oder Grafiken zu den Aufgaben 4 - 6
- **5-Minuten Pitch: Jede Gruppe präsentiert kurz ihre Ergebnisse (5 min maximal) zu Beginn der nächsten Veranstaltung!**