FernUniversität in Hagen

Data Engineering for Data Science

KE2

April 20, 2024

Name:Ricardo Stolzlechner
Matrikelnummer:9463470

---

[1]:
```
# Imports
import pandas as pd
```

Bitte beachten:

- In Code-Zellen soll (bei Bedarf kommentierter) Code eingefügt werden. In Raw-Zellen erwarten wir Antworten im Freitext-Format.
- Fragestellungen sind teilweise bewusst etwas offener formuliert, weil Sie auch im Arbeitsleben nur selten mit ganz spezifischen Anweisungen rechnen können. Es gibt häufig mehrere mögliche Lösungen, dies wird in der Korrektur berücksichtigt.

---

1 (I can't get no) satisfaction

Wir werden in diesem Kurs mit einem Datensatz arbeiten, der verschiedene Informationen über Angestellte einer fiktiven Firma namens *Fiktiva* enthält. Mithilfe des Datensatzes soll die Fragestellung beantwortet werden, ob ein*e Angestellte*r in der Firma zufrieden ist (`satisfied = 1`) oder nicht (`satisfied = 0`). In dieser Kurseinheit werden wir uns mit den Daten vertraut machen und mit einigen Fragen zur Datenqualität auseinandersetzen. Wir gehen für die weiteren Analysen davon aus, dass diese Analyse, die Sie nun durchführen, Anfang März 2021 stattfindet (und somit kurz nach der Erhebung der Daten).

Manchmal werden zur Beantwortung von Fragen weitere Informationen benötigt. Sie können hierfür auch externe Quellen verwenden und bspw. die Datei `metadata.txt` zu Rate ziehen, aber auch alle Informationen aus dem Fiktiva-Newsletter.

1.1 Daten laden

Den Datensatz (*employees_satisfaction.csv*) finden Sie im Moodle. Laden Sie ihn in ein Pandas-Dataframe, so dass die erste Spalte als Index (Row Label) verwendet wird. Geben Sie die ersten 10 Zeilen aus.

[2]:
```
# Ihre Lösung

# daten laden
employee_data = pd.read_csv('../data/fiktiva/employees_satisfaction.csv',␣
 ↪index_col=0)

#head (ersten 10 Zeilen) erzeugen und ausgeben
employee_data_head = employee_data.head(10)
employee_data_head
```

[2]:

	emp_id	age	Dept	education	recruitment_type	job_level	rating	\
0	HR8270	28	HR	PG	Referral	5	2.0	
1	TECH1860	50	Technology	PG	Recruitment Agency	3	5.0	
2	TECH6390	43	Technology	UG	Referral	4	1.0	
3	SAL6191	44	Sales	PG	On-Campus	2	3.0	
4	HR6734	33	HR	UG	Recruitment Agency	2	1.0	
5	PUR7265	40	Purchasing	UG	Walk-in	3	3.0	
6	PUR1466	26	Purchasing	UG	Referral	5	5.0	
7	TECH5426	25	Technology	UG	Recruitment Agency	1	1.0	
8	HR6578	35	HR	PG	Referral	3	4.0	
9	TECH9322	45	Technology	PG	Referral	3	3.0	

	awards	certifications	salary	gender	entry_date	last_raise	satisfied
0	1	0	86750	m	2019-02-01	NaN	1
1	2	1	42419	Male	2017-01-17	NaN	0
2	2	0	65715	f	2012-08-27	NaN	1
3	0	0	29805	f	2017-07-25	NaN	1
4	5	0	29805	m	2019-05-17	NaN	1
5	7	1	42419	m	2004-04-22	0.02	1
6	2	0	86750	m	2019-12-10	NaN	1
7	4	0	24076	Female	2017-03-18	NaN	0
8	0	0	42419	f	2015-04-02	NaN	1
9	9	0	42419	f	2004-03-19	0.07	0

1.2 Erste Datenanalyse

Machen Sie sich mit den Daten vertraut: - Wie viele Zeilen hat der Datensatz? - Wie viele Spalten hat der Datensatz? - Welche Spalten beinhaltet der Datensatz? - Welchen Datentyp haben die einzelnen Spalten? - Ermitteln Sie statistische Kennzeichen der numerischen Spalten. - Gibt es fehlende Werte in den Daten und wenn ja, bei welchen Spalten?

[19]:
```
# Ihre Lösung (nutzen Sie dafür untenstehende Zellen)
```

[3]:
```
# Wie viele Zeilen hat der Datensatz?
print(f'Der Datensatz hat {len(employee_data)} Zeilen')
```

```
Der Datensatz hat 500 Zeilen
```

[4]:
```
# Wie viele Spalten hat der Datensatz?
print(f'Der Datensatz hat {len(employee_data.transpose())} Spalten')
```

```
Der Datensatz hat 14 Spalten
```

[5]:
```
# Welche Spalten beinhaltet der Datensatz?
for col in employee_data.columns:
    print(col)
```

```
emp_id
age
Dept
education
recruitment_type
job_level
rating
awards
certifications
salary
gender
entry_date
last_raise
satisfied
```

[6]:
```
# Welchen Datentyp haben die einzelnen Spalten?
datatypes = employee_data.dtypes
# Ausgabe der Datentypen. Achtung string und Datumsformate werden hier als␣
 ↪object erkannt
# satisfied wird auch als numerischer Wert angegeben (entspricht aber einem␣
 ↪boolschen Wert)
# Evtl. sollte man hier casten bzw. die dtype Option beim einlesen verwenden
datatypes
```

[6]:
```
emp_id              object
age                  int64
Dept                object
education           object
recruitment_type    object
job_level            int64
rating             float64
```

```
awards                int64
certifications        int64
salary                int64
gender               object
entry_date           object
last_raise          float64
satisfied             int64
dtype: object
```

[7]:
```
# Ermitteln Sie statistische Kennzeichen der numerischen Spalten.
# Fehlende Werte werden hier rausgefiltert (eine andere Möglichkeit wäre bspw.
# bei last_raise die Werte mit 0 zu ersetzen)

# Jede Spalte durchgehen
for col in employee_data.columns:
    #typ ermitteln
    column_dtype = employee_data[col].dtype
    if pd.api.types.is_numeric_dtype(column_dtype):
        #wenn es ein numerischer typ ist zuerst daten bereinigen (NaN␣
 ↪rausfiltern, dann statistic berechnen)
        cleaned_data = employee_data.dropna(subset=[col])
        stats = cleaned_data[col].describe()
        #Ausgabe
        print(stats)
        print()
```

```
count    500.000000
mean      39.694000
std        8.477033
min       22.000000
25%       34.000000
50%       39.000000
75%       47.000000
max       56.000000
Name: age, dtype: float64

count    500.000000
mean       3.032000
std        1.423738
min        1.000000
25%        2.000000
50%        3.000000
75%        4.000000
max        5.000000
Name: job_level, dtype: float64

count    471.000000
```

```
mean       3.093418
std        1.423129
min        1.000000
25%        2.000000
50%        3.000000
75%        4.000000
max        5.000000
Name: rating, dtype: float64

count    500.000000
mean       4.570000
std        2.989812
min        0.000000
25%        2.000000
50%        5.000000
75%        7.000000
max       25.000000
Name: awards, dtype: float64

count    500.000000
mean       0.514000
std        0.628167
min        0.000000
25%        0.000000
50%        0.000000
75%        1.000000
max        9.000000
Name: certifications, dtype: float64

count      500.000000
mean     50416.056000
std      23671.392661
min      24076.000000
25%      29805.000000
50%      42419.000000
75%      65715.000000
max      86750.000000
Name: salary, dtype: float64

count    26.000000
mean      0.049231
std       0.029519
min       0.010000
25%       0.020000
50%       0.050000
75%       0.077500
max       0.100000
Name: last_raise, dtype: float64
```

```
count    500.000000
mean       0.714000
std        0.452342
min        0.000000
25%        0.000000
50%        1.000000
75%        1.000000
max        1.000000
Name: satisfied, dtype: float64
```

[8]:
```python
# Gibt es fehlende Werte in den Daten und wenn ja, bei welchen Spalten?

# Alle Spalten durchgehen
for col in employee_data.columns:
    # Fals Eintrag fehlt wird dieser NaN
    if employee_data[col].isna().any():
        print(f"Die Spalte '{col}' hat fehlende Werte")
```

```
Die Spalte 'rating' hat fehlende Werte
Die Spalte 'gender' hat fehlende Werte
Die Spalte 'last_raise' hat fehlende Werte
```

1.3 Datenqualität

Hinweis: Bei den folgenden Fragen geht es um Ihre Einschätzung. Ziel ist es, sich mit dem Thema Datenqualität auseinanderzusetzen und sich mit Fragestellungen in Bezug auf Data Fitness, etc. zu beschäftigen.

Was denken Sie über den Datensatz? Ist dieser geeignet, um die Fragestellung über die Zufriedenheit der Angestellten zu beantworten? Reichen die angegebenen Informationen zur Beantwortung der Fragestellung aus? Wenn nicht, welche Informationen würden Sie noch verwenden? Begründen Sie Ihre Antworten.

Hinweis: Bei der Beurteilung geht es um die Daten selbst, nicht um deren Erhebung. Dazu folgt noch eine Fragestellung.

```
Ihre Einschätzung:
Die Fragestellung lautet: "Wie zufrieden sind die Angestellten?" Problematisch sehe ich
 hierbei, dass das Attribut satisfied offensichtlich als binärer Wert modelliert wurde.
 Das bedeutet, dass die Angestellten nur angeben konnten, ob sie zufrieden sind oder
 nicht. In einem Arbeitsverhältnis gibt es viele Aspekte; wenn nun ein Angestellter in
 einigen Bereichen zufrieden ist und in anderen nicht, muss er sich entscheiden, ob er
 „"zufrieden oder „nicht "zufrieden ist. Besser wäre es hier, eine Skalierung
 einzuführen, beispielsweise das Vergeben von 1-5 Punkten (1 steht für überhaupt nicht
 zufrieden, 5 für voll zufrieden). Das würde die Aussagekraft der Umfrage steigern.
 Weiterhin könnte die Zufriedenheit weiter aufgesplittet werden, wie z. B.
 „Zufriedenheit mit den "Arbeitszeiten, „Zufriedenheit mit der "Arbeitslast,
```

„Zufriedenheit mit der Kommunikation in der "Firma usw. So würde sich ein umfassenderes Bild ergeben.

Für welche weiteren Fragestellungen/Analysen könnte der Datensatz geeignet sein? Zählen Sie mindestens drei sinnvolle Aspekte auf.

Ihre Einschätzung:
Man kann den Datensatz clustern und dann für jeden Cluster die durchschnittliche Zufriedenheit ermitteln, um die einzelnen Gruppen zu ordnen. So erkennt man, in welchen Bereichen die Mitarbeiter am zufriedensten bzw. unzufriedensten sind. Mögliche Cluster sind:

- Dept: Welche Abteilungen haben die zufriedensten Mitarbeiter?
- education: Hat der Ausbildungsgrad einen Einfluss auf die Zufriedenheit?
- salary: Welche Auswirkung hat das Gehalt auf die Zufriedenheit?
- gender: Gibt es Unterschiede in der Zufriedenheit zwischen den Geschlechtern?
- job_level: Steigt die Zufriedenheit mit höherem Joblevel?
- rating: Sind höher bewertete Mitarbeiter zufriedener?

Diese Analysen können Einsichten in die Faktoren liefern, die die Mitarbeiterzufriedenheit beeinflussen. Sie helfen dem Unternehmen, gezielte Maßnahmen zur Verbesserung der Arbeitsumgebung und zur Steigerung der Mitarbeiterzufriedenheit zu entwickeln.

Stellen Sie Überlegungen zur Datenerhebung an. Welche Faktoren könnten die Datenqualität beeinträchtigen? Nennen Sie mindestens zwei Einschränkungen und begründen Sie Ihre Antworten.

Ihre Einschätzung:
1. Offensichtlich war die Abstimmung nicht anonym. Aus den Daten lassen sich persönliche Rückschlüsse ziehen, welcher Mitarbeiter "zufrieden" ist und welcher nicht. Dies kann evtl. über die Mitarbeiter-ID oder über das Gehalt und das Eintrittsdatum erfolgen. Wenn Mitarbeiter befürchten, dass sie Nachteile erleiden, wenn sie mit "nicht zufrieden" abstimmen, könnten sie dazu geneigt sein, stattdessen "zufrieden" anzugeben.
2. Wie bereits erwähnt, ist die Zufriedenheit nicht binär; eine Einteilung in Stufen wäre aussagekräftiger. Die derzeitige binäre Modellierung (zufrieden/unzufrieden) könnte zu einer verzerrten Darstellung der tatsächlichen Mitarbeiterstimmung führen kann.
3. Die Daten wurden nur einmalig erhoben. Wenn die Daten zu einem für die Mitarbeiter besonders stressigen oder untypischen Zeitpunkt erfasst wurden, könnte dies die Ergebnisse verzerren. Beispielsweise könnten Mitarbeiter kurz nach einer Gehaltserhöhung oder während eines umfangreichen Projekts temporär zufriedener oder unzufriedener sein als üblich. Regelmäßige Erhebungen über einen längeren Zeitraum hinweg würden ein realistischeres Bild der durchschnittlichen Mitarbeiterzufriedenheit liefern und saisonale oder ereignisspezifische Schwankungen ausgleichen.

Beurteilen Sie den Datensatz bzgl. der Aspekte zur Datenqualität, die Sie im Kurstext in Kapitel 3 unter ”Assessing Data Integrity” kennengelernt haben. Begründen Sie Ihre Antworten.

[14]:
```
# In dieser Zelle haben Sie Platz für Analysen bzgl. der Datenqualität
# Bewertet werden aber nur Ihre Antworten in der untenstehenden Raw-Zelle, gekennzeichnet mit "Ihre Einschätzung"!
print(employee_data.value_counts('education'))
print()
```

```
print(employee_data.value_counts('recruitment_type'))

print(employee_data.value_counts('gender'))
print(sum(employee_data.value_counts('gender')))
print()
print(employee_data.value_counts('rating'))
print(sum(employee_data.value_counts('rating')))
print()
print(employee_data.value_counts('last_raise'))
print(sum(employee_data.value_counts('last_raise')))
print()
print(employee_data.value_counts('Dept'))
print(sum(employee_data.value_counts('Dept')))
```

```
education
PG    254
UG    246
dtype: int64

recruitment_type
Referral              140
Recruitment Agency    132
Walk-in               116
On-Campus             112
dtype: int64
gender
m         207
f         187
Male       57
Female     46
dtype: int64
497

rating
4.0    111
5.0    100
1.0     89
2.0     89
3.0     82
dtype: int64
471

last_raise
0.05    6
0.01    5
0.02    3
0.04    3
```

```
0.08    3
0.09    3
0.07    2
0.10    1
dtype: int64
26

Dept
Purchasing    114
HR            106
Technology     98
Marketing      95
Sales          87
dtype: int64
500
```

Ihre Einschätzung:
- Is It Timely: Ja, die Daten sind aktuell, da sie, wie vorausgesetzt, Anfang März 2021 ausgewertet wurden.
- Is It Complete: Ja, laut Newsletter gibt es 500 Mitarbeiter, und der Datensatz besteht aus 500 Einträgen.
- Is It Well-Annotated: Teilweise.
    - 'education': Es wird nicht erklärt, was PG und UG bedeuten. Obwohl eine Internet-Recherche ergab, dass UG für "undergraduate" und PG für "postgraduate" steht, wäre es hilfreich, diese Abkürzungen direkt in der metadata.txt zu erläutern.
    - 'recruitment_type': Die verschiedenen Typen des Attributs sollten in den Metadaten beschrieben werden.
    - 'job_level': Die Bedeutung der Stufen 1-5 ist nicht vollständig klar, außer dass Stufe 1 die niedrigste und 5 die höchste ist.
    - 'awards': Es ist unklar, was genau unter "awards" zu verstehen ist. Falls es sich um Ehrungen bzgl. Dienstdauer handelt, könnte diese Information möglicherweise durch das entry_date abgedeckt sein.
    - 'certifications': Was es bedeutet, zertifiziert zu sein, ist aus den Metadaten nicht ersichtlich.
    - 'last_raise': Es ist nicht definiert, was ein leeres Feld bedeutet (möglicherweise, dass es noch nie eine Gehaltserhöhung gab).
- Is It High Volume: Ja, aber mit Einschränkungen aufgrund fehlender Werte bei gender, rating und last_raise.
    - 'gender': Hier fehlen 3 Einträge. Bei der Clusterung nach Geschlecht könnten diese Einträge herausgefiltert oder durch den häufigsten Wert (in diesem Fall 'm') ersetzt werden.
    - 'rating': 29 Felder fehlen. Eine ähnliche Vorgehensweise wie bei gender ist möglich.
    - 'last_raise': Dieser Wert ist nur in 26 Fällen ausgefüllt. Es muss geklärt werden, was ein leerer Wert bedeutet.
- Is It Consisten: Ja, wie aus statistischen Werten der numerischen Felder ersichtlich.
- Is It Multivariate: Ja, es gibt 14 Spalten, wobei awards und entry_date möglicherweise dasselbe aussagen.
- Is It Atomic: Ja, es gibt für jeden Mitarbeiter eindeutige Datenfelder.
- Is It Clear: Abgesehen von den Einträgen PG und UG in der Spalte education ja.
- Is It Dimensionally Structured: Ja, man kann beispielsweise anhand der Abteilung oder des Alters clustern.
- Validity: Die beschriebenen Probleme bei der Art der Datenerhebung könnten die

Validität beeinflussen.
- Reliability: Wie bereits erwähnt, könnte die satisfied-Spalte durch eine 1-5-Skala modelliert werden, um eine zuverlässigere Darstellung der Mitarbeiterzufriedenheit zu ermöglichen.
- Representativeness: Ist gegeben, da jeder Mitarbeiter im Datensatz abgebildet ist.

Stellen Sie sich vor, dieser Datensatz soll dazu verwendet werden, um die Zufriedenheit von Angestellten vorherzusagen und daraus wiederum Rückschlüsse auf das Kündigungswahrscheinlichkeiten zu schließen. Für wie zuverlässig würden Sie entsprechende Wahrscheinlichkeiten halten? Begründen Sie Ihre Antwort.

Ihre Einschätzung:
Ich schätze die Zuverlässigkeit der Vorhersagen auf Basis dieses Datensatzes als moderat ein. Ein wesentliches Problem ist, wie bereits mehrfach erwähnt, dass das Attribut satisfied als binärer Wert modelliert wurde, statt als differenzierte Skala. Diese Darstellung kann die Mitarbeiterzufriedenheit nicht vollständig einfangen, was die Vorhersagegenauigkeit bezüglich der Kündigungswahrscheinlichkeiten beeinträchtigen kann. Eine detailliertere Erfassung der Zufriedenheit in verschiedenen Arbeitsbereichen würde eine präzisere Analyse und damit verlässlichere Vorhersagen ermöglichen.

Ein weiterer kritischer Punkt ist die nicht anonyme Datenerfassung, die möglicherweise zu Verzerrungen führt, da Mitarbeiter aus Furcht vor Nachteilen nicht ihre wahre Zufriedenheit ausdrücken könnten. Dies könnte die Validität der Daten erheblich beeinträchtigen.

Für eine robustere Vorhersage der Kündigungswahrscheinlichkeiten wäre es zudem vorteilhaft, regelmäßige Erhebungen durchzuführen, bspw. in quartalsweisen Abständen. Diese könnten dann direkt mit dem Kündigungsverhalten der Mitarbeiter verknüpft werden, um Trends und Muster über die Zeit zu erkennen und vorherzusagen. Diese regelmäßigen Datenpunkte würden es ermöglichen, die Veränderungen in der Zufriedenheit besser zu verstehen und frühzeitig Indikatoren für mögliche Kündigungen zu identifizieren.

Nehmen Sie an, mithilfe des Datensatzes wäre ein zuverlässiges Modell mit guten Vorhersagewahrscheinlichkeiten trainiert worden, um die Zufriedenheit der Angestellten zu bestimmen. Nun möchte eine andere Firma aus der gleichen Branche dasselbe Modell für ihre Angestellten nutzen. Denken Sie, das Modell würde auch für die andere Firma zuverlässige Vorhersagen treffen? Begründen Sie Ihre Antwort.

Ihre Einschätzung:
Es gibt viele Bereiche, in denen sich Unternehmen unterscheiden, und das beschränkt sich nicht nur auf die Branche. Beispiele hierfür sind die Firmenkultur und der Umgang der Führungspersonen mit ihren Mitarbeitern. Deshalb glaube ich nicht, dass sich die Daten in der vorliegenden Form, die lediglich ein binäres Zufriedenheitsattribut verwenden, eignen. Vor einer Anwendung müsste vor allem eine Validierung des Modells erfolgen, und es sollte gegebenenfalls angepasst werden.

Überlegen Sie, welche Vorverarbeitungsschritte für den Datensatz notwendig wären. Nennen Sie mindestens drei sinnvolle Vorverarbeitungsschritte.

Ihre Einschätzung:
1. Gender anpassen: Die Werte in der Spalte 'gender' sind nicht einheitlich und umfassen 'm', 'Male', 'f' und 'Female', wobei 'f' und 'Female' sowie 'm' und 'Male' dasselbe ausdrücken. Diese Werte sollten vereinheitlicht werden, indem 'm' zu 'Male' und 'f' zu 'Female' geändert wird.

2. Fehlende Werte auffüllen: Einige Spalten weisen fehlende Werte auf:
   - 'gender': Hier kann man entweder die fehlenden Werte mit dem am häufigsten vorkommenden Wert ersetzen oder die Einträge herausfiltern.
   - 'rating': Dasselbe Vorgehen wie bei 'gender'.
   - 'last_raise': Es muss festgestellt werden, was ein leerer Wert bedeutet. Falls ein leerer Wert bedeutet, dass es noch nie eine Gehaltserhöhung gab, könnte man diese mit '0' ersetzen.
3. Umwandlung in korrekten Datentyp: Einige Spalten enthalten Strings und Datumsformate, die in pandas als 'object' typisiert sind. Diese Werte sollten in die entsprechenden Datentypen umgewandelt werden, beispielsweise Datumsangaben in 'datetime' und strings in 'str'.