|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Gruppe | Studierende 1 | | Studierende 2 | |
|  | Kürzel | Name, Vorname | Kürzel | Name, Vorname |
|  |  |  |  |
| Studierende 3 | | Studierende 4 | |
| Kürzel | Name, Vorname | Kürzel | Name, Vorname |
|  |  |  |  |

# Thema

Datenauswertung am Beispiel von Business-Anwendungsfällen

# Lernziele

1. Sie haben einen Einblick in die Vorgehensmodelle der Softwareentwicklung im Team.

2. Sie können in einer größeren Gruppenarbeit gemeinsam eine Spezifikation erarbeiten.

3. Sie können Anforderungen erheben und mit Ihnen Umgehen.

4. Sie können eine Datenauswertung durchführen.

5. Sie können ihre Ergebnisse im Plenum präsentieren.

# Hinweise

* Bitte tragen Sie die fehlenden Daten im Kopfbereich ein.
* Bitte verwenden Sie vorzugsweise dieses Dokument für die Inhalte Ihrer Abschlusspräsentation. Ansonsten verwenden Sie bitte ihr bevorzugtes Präsentationsformat,  
  aber achten Sie bitte darauf, dass Sie die Datei(en) als PDF exportieren können. Bei mehreren Dateien packen Sie bitte alles in ein ZIP-Archiv.
* Bitte verwenden Sie folgende Dateinamen-Syntax: E1\_2021ss\_Gruppe  
  Ändern Sie Gruppe entsprechend.
* Bitte verwenden Sie für den Upload das Dateiformat PDF. Sollten Sie mehrere Dateien abgeben, dann packen Sie alles in ein ZIP-Archiv.
* Die Abgabe erfolgt fristgerecht durch den Gruppenverantwortlichen per Upload in ILIAS.
* Die Bewertung berücksichtigt teilweise korrekte Lösungen. Deshalb werden die Punkte summiert und letztlich durch 100 geteilt, um auf die endgültige Punktzahl zu kommen,  
  die Sie als Startwert für die Klausur anrechnen können.

# Aufgabenstellung

Die Aufgabenstellung finden Sie auf GitHub im Verzeichnis p in der Readme.md Datei, siehe [Link](https://github.com/hka-mmv/dscb230-exercise/tree/main/p#readme).

[Aufgabe: Datenauswertung am Beispiel von Airbnb NYC (300 Punkte) 2](#_Toc74923379)

## Aufgabe: Datenauswertung am Beispiel von Airbnb NYC (300 Punkte)

### Entwicklungsschritte

Zuerst bearbeiteten wir den Punkt Business Understanding. Hierzu schauten wir uns das Data-Set genauer an, indem wir auf Kaggle die zur Verfügung gestellten Metadaten sowie die allgemeinen Informationen des Sets durchgelesen haben. Weiterhin haben wir die csv-Datei in einem Jupyter-Notebook mittels Pandas in ein DataFrame geladen um ein genaueres Verständnis über die Daten zu erhalten. Hierfür wendeten wir die Methoden dtypes() an um zu schauen, welche Spalte welchen Datentyp hat. Hier viel uns auf, das beispielsweise die Spalte „last\_review“ von Typ „Object“ ist. Daher parsten wir im weiteren Schritt des „Data Cleaning“ diese Spalte als ein Datetime-Objekt. Zur Datenanalyse ließen wir uns zudem über die Methode describe() einen weiteren Überblick über die Daten geben. Mittels dieser Methode konnten wir beispielsweise sehen, dass der niedrigste Preis bei 0 Dollar liegt und der höchste bei 10000 Dollar. Außerdem haben die Werte der Spalte „price“ eine Standardabweichung von ca. 240 Dollar. Bei einem ersten Überblick über die Daten ist uns ebenfalls aufgefallen, dass in der Spalte „host\_name“ oftmals nur Vornamen beziehungsweise ein Unternehmensname vorkamen. Manchmal auch nur ein Buchstaben, was einem nicht viel weiter bringt.

Im anschließenden Schritt, dem „Data Cleaning“, bereinigten wir die Daten. Hierzu summierten wir die Null-Werte der einzelnen Spalten mittels df.isna().sum() auf um zu schauen, welche Spalten Null-Werte enthalten und entsprechend wie viele. Anschließend überlegten wir, wie wir mit den entsprechenden Null-Werten umgehen. Wir entschieden uns die Zeilen der Spalte „name“, bei denen Null-Werte enthalten waren, komplett aus dem DataFrame zu entfernen. Dies waren 16 Zeilen. Wir entschieden uns für diesen Schritt, da wir uns überlegten, dass es keinen Sinn macht mit einem Airbnb-Objekt zu arbeiten, dass keinen Namen hat. Uns ist weiterhin aber auch bewusst, dass dies zu einer Verfälschung von anderen Auswertungen kommen kann, beispielsweise in Bezug auf die Anzahl der Reviews. Jedoch fallen 16 Zeilen bei einem Data-Set mit insgesamt 48895 Zeilen nicht so stark ins Gewicht. Anders sind wir mit den Nullwerten der Spalte „reviews\_per\_month“ umgegangen. Wir sind hierbei davon ausgegangen, dass, wenn diese Zahl fehlt, dort eine Null stehen müsste. Daher füllten wir in dieser Spalte die entsprechenden Zeilen mit Nullen auf. Hierfür nutzten wir die Methode fillna(0). Auch in der Spalte „last\_review“ gibt es Null-Werte. Hier entschieden wir uns, diese beizubehalten, da diese Spalte Datumswerte enthält und wir keinen plausiblen Standardwert gefunden haben. Ebenfalls in der Spalte „host\_name“ behielten wir Null-Werte (hier insgesamt 21 Null-Werte), da wir es an der Stelle nicht so wichtig fanden, ob der Name des Vermieters angegeben ist oder nicht. Das bereinigte Data-Set schrieben wir dann in eine neue csv-Datei (AB\_NYC\_2019\_CLEAN.csv).

### Ergebnisse

Ihr Text

### Schwierigkeiten und Lösungsansätze

Ihr Text

### Live-Demo

[ ] Ja.

[ ] Nein. Begründung: