|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Gruppe | Studierende 1 | | Studierende 2 | |
| **India** | Kürzel | Name, Vorname | Kürzel | Name, Vorname |
| **Bial1012** | **Bickel, Alina** | **Duma1024** | **Dürr, Maximilian** |
| Studierende 3 | | Studierende 4 | |
| Kürzel | Name, Vorname | Kürzel | Name, Vorname |
| **Kiya1013** | **Kircher, Yannick** | **Jami1013** | **Jäckle, Michael** |

# 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).

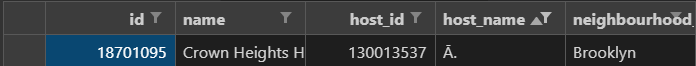
[Aufgabe: Datenauswertung am Beispiel von Airbnb NYC (300 Punkte) 2](#_Toc74923379)

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

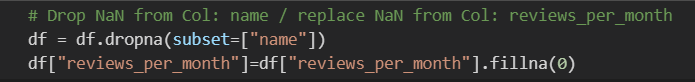
## Entwicklungsschritte

### Data Understanding

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. Bei einer ersten Sichtung der Daten mittels eines Excel-Viewers in VS Code fiel weiterhin auf, dass der „host\_name“ zum Teil auch nur aus einem Buchstaben bestand, was einem nicht viel weiter bringt.



### Data Cleaning

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) um darauf Auswertungen machen zu können und diese auch zu visualisieren.

Ein Bild, das Text enthält.

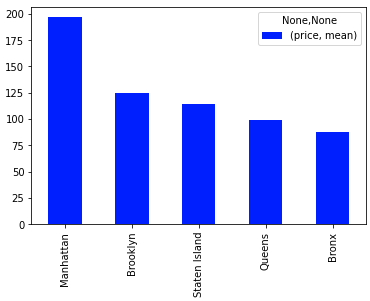
Automatisch generierte Beschreibung

### Datenvisualisierung

##### Group By Abfragen

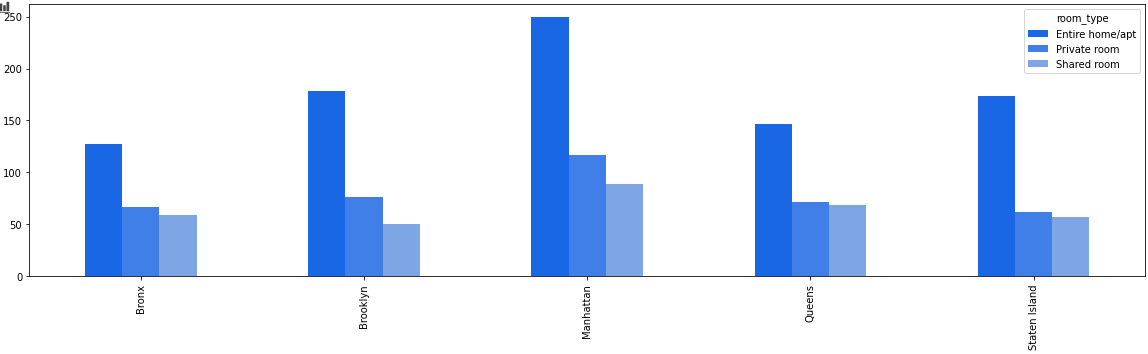
Hier hatten wir das Ziel die Daten sinnvoll, visuell aufzubereiten. Angefangen haben wir mit vier einfachen „Group by“ Abfragen. Diese haben wir dann in Jupyter Lab ausgeben lassen, um Einsicht in die Daten zu erhalten.

**z.B.:**

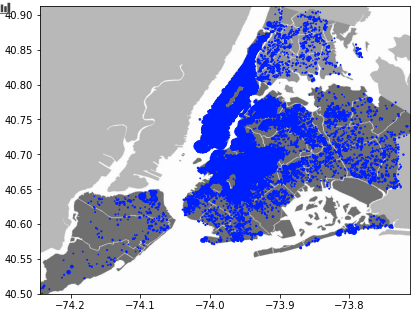
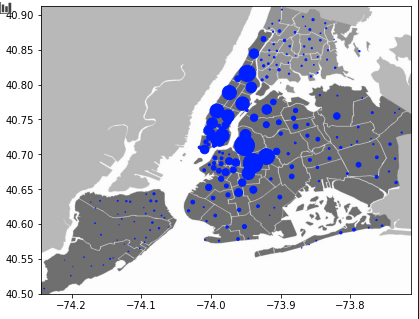
Berechnen des durchschnittlichen Preises und die Anzahl der Preise, die für die Berechnung einbezogen wurden gruppiert nach den Spalten 'neighbourhood\_group' und 'neighbourhood'.

##### Histogramme

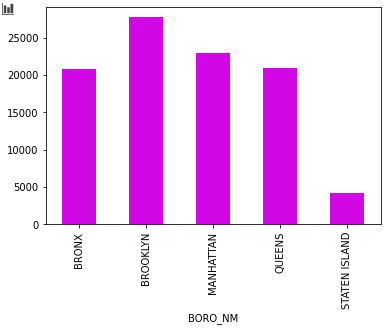
Weiterhin haben wir Histogramme erstellt, auch wieder über „Group by“ Abfragen auf dem Pandas Dataframe. Hier haben wir zum Beispiel den Durchschnittlichen Preis und Anzahl Listings Pro Stadtteil geplottet. Allerdings haben wir auch den Zimmertyp Analysier und uns den Durchschnittlichen Preis des jeweiligen Zimmertyps pro Stadtteil Ausgegeben.



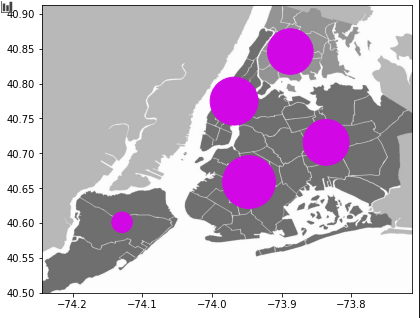
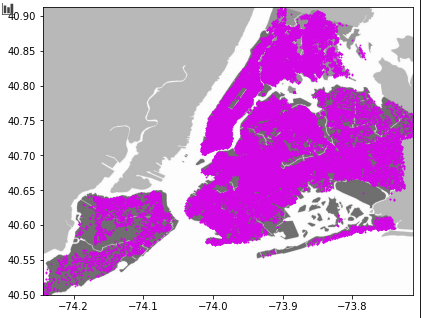
##### Scatter-Maps

Zunächst haben wir uns alle Listings auf der Karte von New York anzeigen lassen. Dies war möglich da wir zu jedem Listing die Längen und Breitengrade gegeben hatten. Um die Dichte der Listings besser erkennen zu können haben dann diese gruppiert nach der Spalte „neighbourhood“.

##### Histogramme (Crime)

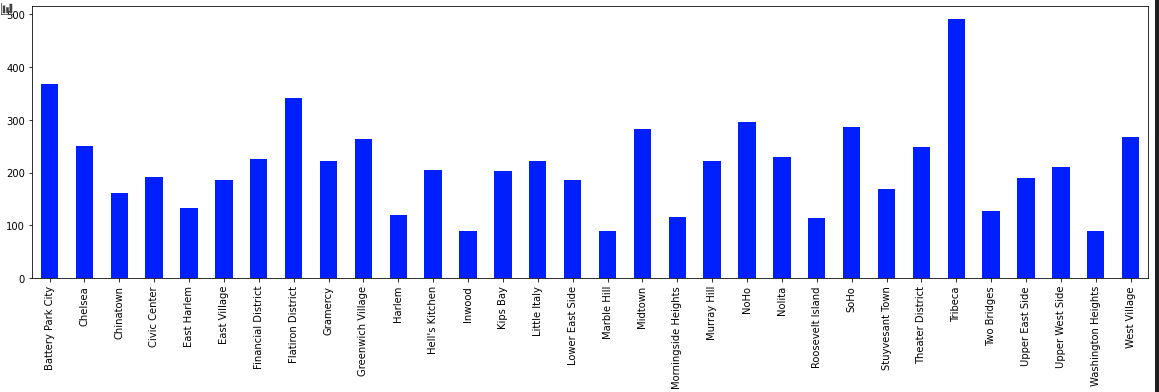
Hier haben wir mit dem „NYPD\_Complaint\_Map\_\_Year\_to\_Date“ Data Set gearbeitet. Das Vorgehen war wie beider normalen Analyse, die Crimes wurden gruppier nach Stadtteil ausgegeben.

##### Scatter-Maps (Crime)

Weiterhin haben alle Complaints auf der Karte von New York angezeigt, hier waren uns auch wider die Koordinaten gegeben.

## Ergebnisse

Anhand der „Group By“-Abfragen sowie den Histogrammen und Scatter Maps konnten wir einige Auswertungen auf dem Datensatz durchführen. So sahen wir mittels einer „Group By“-Abfrage, dass in Staten Island zwei recht teure Airbnb-Objekte liegen, mit 800€ und 700€ Pro Nacht. Weiter fanden wir heraus, dass die meisten Vermietungsobjekte vom Typ „Entire home/apt“ sind. Insgesamt wurden 25402 Werte dieser Kategorie gefunden. Es ist allerdings nicht verwunderlich, dass dieses auch mit einem durchschnittlichen Preis von rund 211$ das teuerste Objekte zum Mieten ist. Wobei die Werte von 0$ bis 10000$ pro Nacht reichen. Auch ist hier die Standardabweichung mit rund 284$ recht groß. Danach folgt der Typ „Private room“ mit einem mittleren Preis von ca. 89$. Diese Kategorie ist als zweithäufigsten im Datensatz vorhanden. Auch hier liegt die Spannweite zwischen 0$ und 10000$. Das billigste Airbnb-Objekt mit durchschnittlich 70$ ist vom Typ „Shared room“. Hier beträgt der maximale Preis 1800$. Mittels eines Histogramms wurde deutlich, in welchen „neighbourhood groups“ die meisten Entire home/apt liegen. Hier war deutlich zu erkennen, dass dies in Manhattan der Fall ist. Weiterhin fanden wir heraus, dass in Manhattan die meisten Vermietungsobjekte vorliegen und mit einem deutlichen Abstand hierzu am wenigsten in Staten Island. Durschnittlich ist eine Mietung eines Airbnb-Objektes in Manhattan am teuersten und in Bronx am billigsten.

Da Manhattan in allen Kategorien am Teuersten ist und auch am meisten Listing hat haben wir die Nachbarschaften in Manhattan mit einem weiteren Histogramm analysiert. Hierbei ergab sich, dass der Stadtteil „Tribeca“ im Durchschnitt am teuersten ist.

Bei der Auswertung der Karten haben wir uns zunächst die Bevölkerung New Yorks angesehen und festgestellt das in Manhattan, wo am meisten Listings liegen auch am meisten Menschen wohnen. Die Anzahl an Verbrechen ist in der Bronx, Brooklyn, Manhattan, Queens etwa gleich(um die 21 Tausend) In Statten Island Ist die Anzahl wie erwartet deutlich geringer.(etw. 4 Tausend)

## Schwierigkeiten und Lösungsansätze

Bei der Bereinigung des Datensatzes gab es Probleme, da wir uns nicht sicher waren wie wir mit den Null/Nan Werten umgehen sollen. Wir haben uns dann dazu entschieden nur ein par spalten zu bereinigen da es bei anderen, als nicht notwendig angesehen haben.

Die Karten Analyse stellt eine weiter Schwierigkeit da. Das bereitgestellt Bild von New York hat nicht mit den Koordinaten des Datensatzes übereingestimmt. Dieses Herausforderung wurde durch überprüfen und das zuschneiden des Bildes gelöst.

### Live-Demo

[x ] Ja.

[ ] Nein. Begründung: