Project reporting

0. Offene Punkte
1. Alle 3 Regionen importieren
2. Kategorie 4 abschliessen
3. Bei allen 3 regionen correlation matrix der Variablen vergleichen -> Die mit einer Correlation filtern
4. Fragen wie die Datenquelle angegeben werden soll
5. Dozent fragen ob 2 oder 3 Methoden (Descriptive, Predictive, Prescriptive) umgesetzt werden muss

I. 1. Abstrakt

Document Abstract [1pt]

- Does it provide a comprehensible glipse of the project? [0.5pt]
- Is the outcome teased? [0.5pt]

II. 2. Einleitung (mit Forschungsfrage (d.h. Geschäftsfrage) am Ende)

Questions [2.5pt] Cedric

- Are all BI topics (descriptive, predictiv and prescriptive) covered in the questions? [0.5pt]
- Is the formulation close enough? [1pt]
- Is there a reasonable complexity of the questions (i.e. the BI problems)? [1pt]

Hypothesis [1pt] Are (is) there (a) global hypothesis and motivation(s) to prove? [1pt]

In der heutigen, schnelllebigen Welt des Online-Tourismus spielen Plattformen wie Airbnb eine zentrale Rolle bei der Art und Weise, wie Menschen reisen und Unterkünfte buchen. Airbnb bietet eine Vielzahl von Unterkünften an, von einfachen Zimmern bis hin zu luxuriösen Villen. So vielfältig wie das Angebot sind auch die Vorlieben und Erwartungen der Gäste. Vor diesem Hintergrund stellt sich die Forschungsfrage: ***Welche Eigenschaften einer Airbnb-Unterkunft ziehen Gäste an und ermöglichen es, einen höheren Preis pro Apparment zu erzielen?***

Die Bedeutung dieser Frage ergibt sich aus dem zunehmenden Wettbewerb unter den Gastgebern, die bestrebt sind, ihre Unterkünfte attraktiv zu gestalten und gleichzeitig optimale Preise zu erzielen. Durch die Analyse von Daten aus den Regionen Zürich, Genénve und Vaud versucht diese Arbeit, die Schlüsselelemente zu identifizieren, die die Attraktivität und die Preisgestaltung beeinflussen. Neben der geografischen Lage, der Art der Unterkunft, der Anzahl der Zimmer und Bäder sowie der Verfügbarkeit von Annehmlichkeiten werden auch Faktoren wie die Erfahrung und das Profil des Gastgebers analysiert. Diese Untersuchung zielt darauf ab, die prädiktiven Merkmale einer Unterkunft zu bestimmen, die als Indikatoren für höhere Preise fungieren, und präskriptive Empfehlungen für Gastgeber zu entwickeln, um die Attraktivität ihrer Unterkünfte zu steigern.

Durch die Kombination beschreibender, prädiktiver und präskriptiver Analysen wird die Studie ein tiefgreifendes Verständnis dafür vermitteln, wie verschiedene Faktoren zusammenwirken, um den wahrgenommenen Wert einer Unterkunft aus Sicht des Gastes zu erhöhen. Die Hypothese, dass Unterkünfte mit modernen Annehmlichkeiten, hervorragender Lage und hohen Bewertungen signifikant höhere Preise erzielen, wird überprüft, um Gastgebern und Airbnb wertvolle Einblicke für die Optimierung von Angeboten und Plattform zu bieten.

III. 3. Datenquelle (mit Angaben zu Quelle, Qualität und Bereinigungsschritten der Daten)

Data set [2.5pt] -> Yes Jovan

Comprises the data set a resonable amount of observations to answer the BI questions? [1pt] -> Yes

Bring the additional data set new information to the basic data set? [1pt] -> Yes

Are all data set accessible linked in documentation? [0.5pt] -> Yes

Die Daten für diese Analyse stammen von der Inside Airbnb Organisation (https://insideairbnb.com/get-the-data/), die sich dafür einsetzt, ihre Gemeinden vor den negativen Auswirkungen von Kurzzeitvermietungen zu schützen. Diese Organisation sammelt und veröffentlicht regelmässig aktualisierte Datensätze, die aus öffentlich verfügbaren Informationen auf der Airbnb-Website stammen. Diese Datensätze würden wir als Vertrauenswürdig einstufen.

Die extrahierten Datensätze umfassen Informationen aus drei bedeutenden Regionen in der Schweiz: Zürich (27. Dezember 2023), Genénve (27. Dezember 2023) und Vaud (10. März 2024).

Die Daten umfassen verschiedene Dateien für jede Stadt, wobei für die Analyse hauptsächlich das “listings_long.csv”-File verwendet wird, da es für die Geschäftsfragen relevant ist. Die Qualität der Daten in diesem File ist insgesamt sehr hoch, mit wenigen leeren Feldern und einer konsistenten Struktur innerhalb der Spalten.

Es werden insgesamt 3 Datensätze verwendet, die dieselbe Struktur aufweisen, jedoch aus drei verschiedenen Regionen stammen, die wir hier analysieren möchten.

Einige Spalten, wie “description”, “neighborhood_overview”, “host_neighborhood”, “neighborhood”, und die Beschreibung der Liegenschaft, wie “bathrooms” und “bedrooms”, weisen eine beträchtliche Anzahl leerer Felder auf. Es wird vermutet, dass diese Felder optional für die Gastgeber sind und daher nicht immer ausgefüllt werden. Ebenso fehlt bei einigen Einträgen der Preis, was eine Analyse erfordert, um mögliche Korrelationen mit anderen Feldern, wie dem ersten Review, zu identifizieren.

Trotz dieser kleinen Unregelmässigkeiten ist die Datenqualität insgesamt hoch, und die

Beschreibung der Datenfelder wird durch das Data Dictionary (https://docs.google.com/spreadsheets/d/1iWCNJcSutYqpULSQHlNyGInUvHg2BoUGoNRIGa6Szc4/edit#gid=1322284596) gut unterstützt.

Datenbereinigungsschritte:

1. **Entfernung irrelevanter oder leerer Spalten:** Vor der Analyse wurden alle Spalten entfernt, die für die Fragestellungen nicht relevant sind oder leere Felder enthalten.
2. **Überprüfung der Einheitlichkeit und Konsistenz der Werte:** Die verbleibenden Spalten wurden auf Einheitlichkeit der Werte und Konsistenz der "N/A"-Kennzeichnungen überprüft, um sicherzustellen, dass die Daten konsistent und interpretierbar sind.
3. **Analyse von Einträgen ohne Preisangabe:** Einige Einträge weisen keine Preisangabe auf, was eine Analyse erfordert, um mögliche Korrelationen mit anderen Feldern, wie dem ersten Review, zu identifizieren. Je nach Ergebnis dieser Analyse könnten Einträge ohne Preisangabe entfernt oder anderweitig behandelt werden, um die Datenintegrität zu gewährleisten.

IV. 4. Datenqualität (Analyse im Hinblick auf Datenqualitätsaspekte)

ETL [3pt] Jovan

How qualitative (detailed) is the ETL/EDA process done? [1pt]

What are the findings and how are they discussed? [1pt]

Was the data set cleaned with reasonable approaches? [1pt]

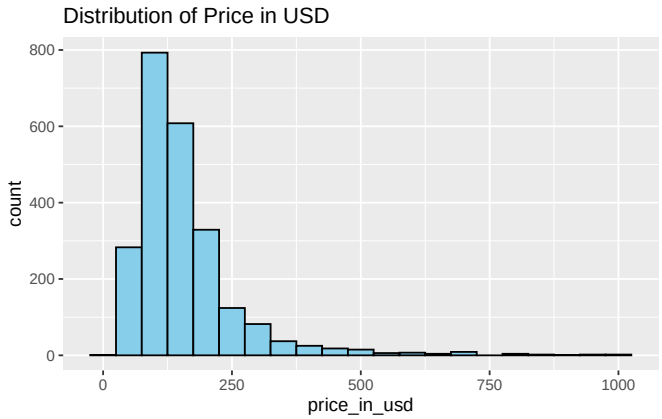
Nachdem wir die Zusammenfassung der Daten betrachtet und eine detaillierte Analyse der Dataframes durchgeführt haben, haben wir die folgenden Datenanpassungen vorgenommen:

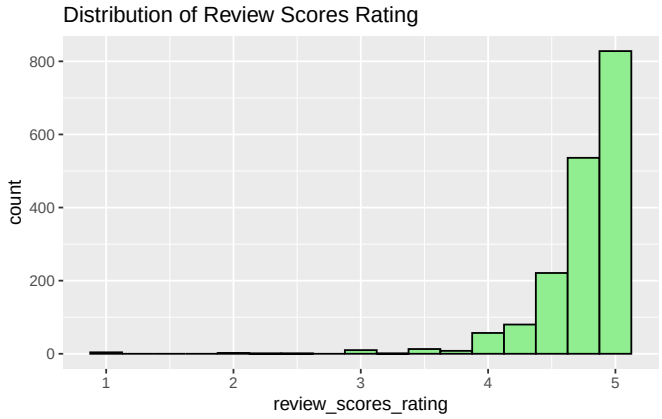
1. Konvertiere **`host_response_rate`** von Zeichenfolge (chr) in Ganzzahl (integer), wobei "N/A" durch NA ersetzt wird. Entferne das Prozentzeichen (%) und benenne die Spalte in **`host_response_rate_in_%`** um.
2. Konvertiere **`host_acceptance_rate`** von Zeichenfolge (chr) in Ganzzahl (integer), wobei "N/A" durch NA ersetzt wird. Entferne das Prozentzeichen (%) und benenne die Spalte in **`host_acceptance_rate_in_%`** um.
3. Konvertiere **`price`** von Zeichenfolge (chr) in Dezimalzahl (double), wobei das Dollarzeichen ($) entfernt wird. Benenne die Spalte in **`price_in_$`** um.
4. Lösche die folgenden Spalten: **`description`**, **`neighborhood_overview`**, **`host_location`**, **`host_about`**, **`host_neighbourhood`**, **`host_verifications`**, **`neighbourhood`**, **`neighbourhood_cleansed`**, **`bathrooms`**, **`bedrooms`**, **`amenities`**, **`calendar_updated`**, **`license`**.
5. Analysiere fehlende Werte (NA) oder leere Felder und ersetze sie gegebenenfalls.

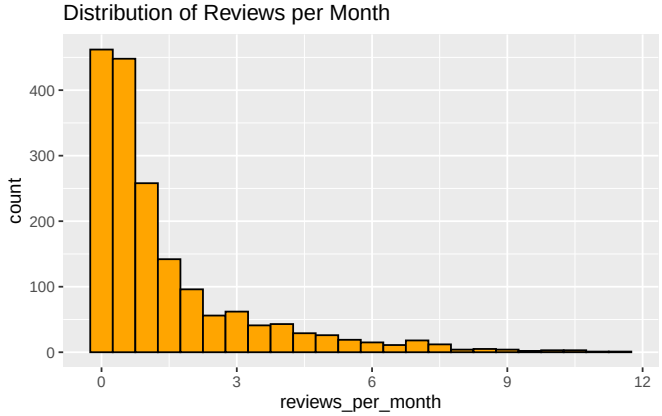
```
 [1] "id"
 [2] "listing_url"
 [3] "scrape_id"
 [4] "last_scraped"
 [5] "source"
 [6] "name"
 [7] "picture_url"
 [8] "host_id"
 [9] "host_url"
[10] "host_name"
[11] "host_since"
[12] "host_response_time"
[13] "host_is_superhost"
[14] "host_thumbnail_url"
[15] "host_picture_url"
[16] "host_listings_count"
[17] "host_total_listings_count"
[18] "host_has_profile_pic"
[19] "host_identity_verified"
[20] "neighbourhood_group_cleansed"
[21] "latitude"
[22] "longitude"
[23] "property_type"
[24] "room_type"
[25] "accommodates"
[26] "bathrooms_text"
[27] "beds"
[28] "minimum_nights"
[29] "maximum_nights"
[30] "minimum_minimum_nights"
[31] "maximum_minimum_nights"
[32] "minimum_maximum_nights"
[33] "maximum_maximum_nights"
[34] "minimum_nights_avg_ntm"
[35] "maximum_nights_avg_ntm"
[36] "has_availability"
[37] "availability_30"
[38] "availability_60"
[39] "availability_90"
[40] "availability_365"
[41] "calendar_last_scraped"
[42] "number_of_reviews"
[43] "number_of_reviews_ltm"
[44] "number_of_reviews_l30d"
[45] "first_review"
[46] "last_review"
[47] "review_scores_rating"
[48] "review_scores_accuracy"
[49] "review_scores_cleanliness"
[50] "review_scores_checkin"
[51] "review_scores_communication"
[52] "review_scores_location"
[53] "review_scores_value"
[54] "instant_bookable"
[55] "calculated_host_listings_count"
[56] "calculated_host_listings_count_entire_homes"
[57] "calculated_host_listings_count_private_rooms
[58] "calculated_host_listings_count_shared_rooms"
[59] "reviews_per_month"
[60] "host_response_rate_in_percent"
[61] "host_acceptance_rate_in_percent"
```

```
[62] "price_in_usd"
```

Um die Daten nach der Bereinigung zu überprüfen und einen Überblick zu erhalten, führen wir eine standardmässige Datenanalysen durch am Beispiel **`df_zuerich_cleanead`** .





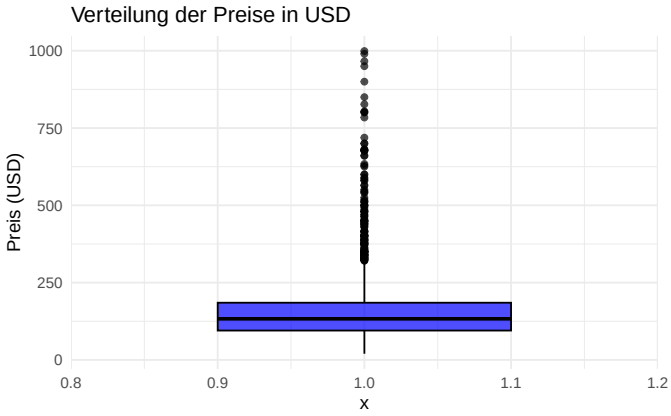


V. 5. Datenanalyse (Informationen zur Datenstruktur, Organisation und zu den für die Analyse verwendeten Methoden)

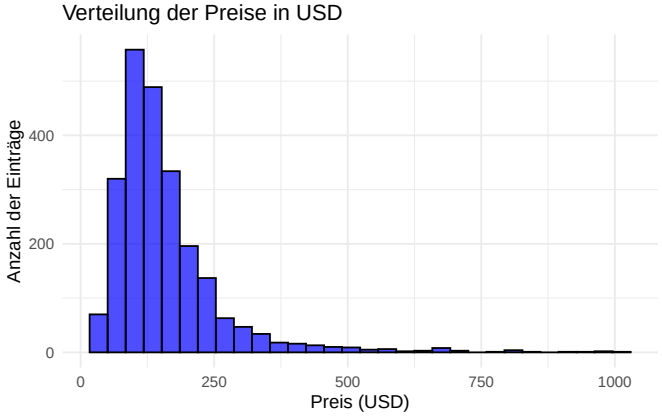
Data analysis [2pt] Cedric / Jovan
Are data attributes clear and obvious discussed? [1pt]
Are adequate EDA methodologies used for data analysis? [1pt]

Die bereinigten Datensets der verschiedenen Orten sind alle gleich aufgebaut. Da wir den Preis der einzelnen Airbnb Apparment anschauen möchten ist dies unser wichtigster Wert:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   20.0     95.0    133.0   157.8   185.0   999.0
```




```
List of 3
 $ axis.title.x: list()
  ..- attr(*, "class")= chr [1:2] "element_blank"
 $ axis.text.x : list()
  ..- attr(*, "class")= chr [1:2] "element_blank"
 $ axis.ticks.x: list()
  ..- attr(*, "class")= chr [1:2] "element_blank"
 - attr(*, "class")= chr [1:2] "theme" "gg"
 - attr(*, "complete")= logi FALSE
 - attr(*, "validate")= logi TRUE
```



Nun ist stellt sich die Frage welche anderen Eigenschaften die grösste Auswirkung auf den Preis haben. Dazu gilt es herauszufinden wie die Korrelationen zwischen dem Preis pro USD und den anderen Attributen sind:

```
                                id
                      -0.051733794
                         scrape_id
                                NA
                      last_scraped
                       0.323817084
                           host_id
                      -0.034408421
                        host_since
                      -0.075722239
                host_listings_count
                      -0.048121517
           host_total_listings_count
                      -0.030748894
```

```
                    latitude              calculated_host_listings_count
                -0.096602270                                  -0.034790038
                   longitude calculated_host_listings_count_entire_homes
                 0.048817534                                  -0.026939926
              accommodatescalculated_host_listings_count_private_rooms
                 0.503562791                                  -0.116312035
                        beds calculated_host_listings_count_shared_rooms
                 0.419758800                                  -0.026939558
               minimum_nights                            reviews_per_month
                -0.100413698                                  -0.022824153
               maximum_nights              host_response_rate_in_percent
                 0.078865097                                   0.030119785
     minimum_minimum_nights            host_acceptance_rate_in_percent
                -0.045297000                                   0.042059103
     maximum_minimum_nights                                 price_in_usd
                -0.024775869                                   1.000000000
     minimum_maximum_nights
                 0.100345512
     maximum_maximum_nights
                 0.041970938
     minimum_nights_avg_ntm
                -0.022747969
     maximum_nights_avg_ntm
                 0.042085419
            availability_30
                 0.103389623
            availability_60
                 0.112909276
            availability_90
                 0.110224160
           availability_365
                 0.130285779
      calendar_last_scraped
                 0.323817084
          number_of_reviews
                 0.003813291
       number_of_reviews_ltm
                -0.016722222
     number_of_reviews_l30d
                -0.021028955
                first_review
                -0.053134782
                 last_review
                 0.039951229
        review_scores_rating
                 0.057822255
     review_scores_accuracy
                 0.023426798
   review_scores_cleanliness
                 0.125133364
      review_scores_checkin
                 0.020947515
review_scores_communication
                 0.006917651
     review_scores_location
                 0.091844483
          review_scores_value
                -0.008147618
```

Die Korrelationen der verschiedenen Variablen mit dem Preis (**price_in_usd**) im Datensatz können wichtige Einsichten bieten, welche Faktoren den Preis beeinflussen. Hier ist eine Analyse der signifikanten positiven und negativen Korrelationen:

A. Positive Korrelationen:

1. **accommodates** (0.449): Es besteht eine moderate positive Korrelation zwischen der Anzahl der Gäste, die eine Unterkunft aufnehmen kann, und dem Preis. Dies deutet darauf hin, dass grössere Unterkünfte, die mehr Gäste beherbergen können, in der Regel teurer sind.
2. **beds** (0.329): Eine ähnliche positive Korrelation gibt es zwischen der Anzahl der Betten und dem Preis, was darauf hindeutet, dass mehr Betten oft höhere Preise bedeuten, was auch mit der Grösse der Unterkunft zusammenhängen kann.
3. **availability_365** (0.157): Längere Verfügbarkeit im Laufe eines Jahres korreliert leicht positiv mit höheren Preisen. Dies könnte bedeuten, dass Unterkünfte, die seltener verfügbar sind, zu höheren Preisen angeboten werden.
4. **last_scraped** (0.308), **calendar_last_scraped** (0.308): Diese Korrelationen deuten darauf hin, dass die Zeitpunkte der Datenerfassung mit den Preisänder

B. Negative Korrelationen:

1. **latitude** (-0.097): Es gibt eine leichte negative Korrelation zwischen der geographischen Breite und dem Preis. Je weiter nördlich die Unterkunft liegt, desto geringer könnte der Preis sein, was auf regionale Preisunterschiede in der Stadt oder der Umgebung hinweisen könnte.
2. **minimum_nights** (-0.053), **first_review** (-0.053): Längere Mindestaufenthalte und frühere erste Bewertungen korrelieren leicht negativ mit dem Preis. Dies könnte darauf hinweisen, dass preisgünstigere Unterkünfte möglicherweise längere Aufenthalte erfordern oder schon länger auf dem Markt sind.
3. **calculated_host_listings_count_private_rooms** (-0.103): Eine höhere Anzahl von Inseraten, die private

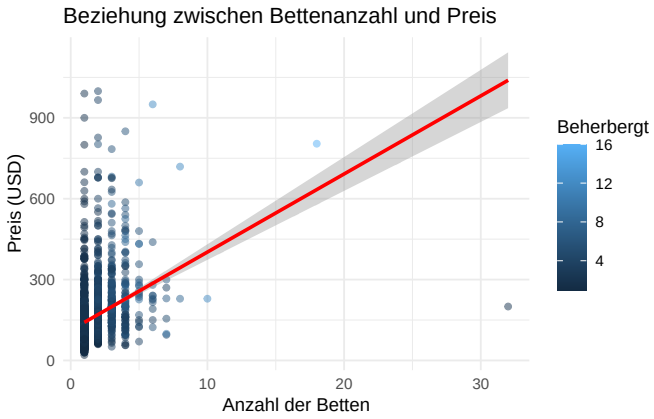
Zimmer betreffen, korreliert leicht negativ mit dem Preis, was darauf hinweisen könnte, dass Gastgeber mit mehreren Einträgen möglicherweise günstigere Preise anbieten, um wettbewerbsfähig zu bleiben.

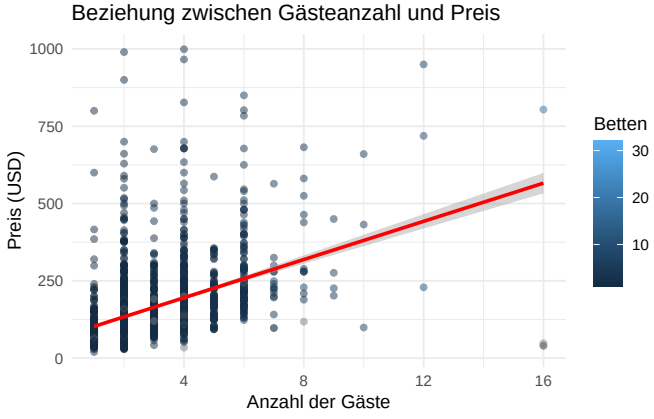
C. *Interpretation*

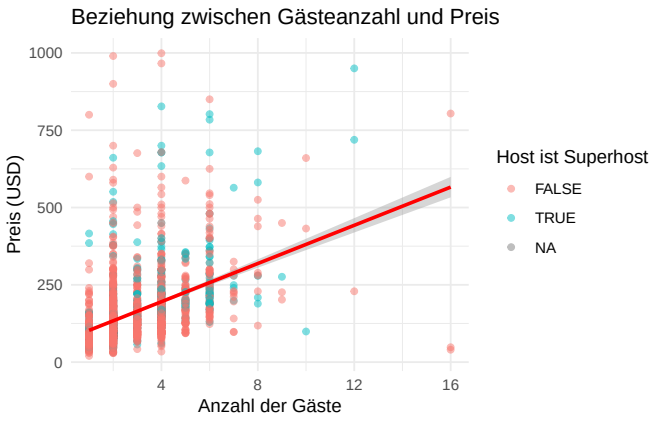
- **Hohe positive Korrelationen** zeigen an, dass mit zunehmender Kapazität und Verfügbarkeit der Unterkünfte der Preis steigt. Dies reflektiert die Marktlogik, dass grössere und häufiger verfügbare Unterkünfte als wertvoller angesehen werden.
- **Negative Korrelationen** deuten darauf hin, dass bestimmte Faktoren wie die geografische Lage (nördlicher) oder die Politik längerer Mindestaufenthalte die Preise senken können. Dies könnte für Gäste attraktiv sein, die längerfristige Aufenthalte suchen oder flexibel in der Wahl der Region sind.

Diese Korrelationen sollten jedoch vorsichtig interpretiert werden, da Korrelation nicht gleich Kausalität ist. Andere verborgene Variablen könnten ebenfalls eine Rolle spielen, und die Effekte könnten durch spezifische Marktbedingungen oder andere nicht berücksichtigte Faktoren beeinflusst werden.

Um diese Korrelationen auch noch grafisch aufzuzeigen mit einer linearen Regression.



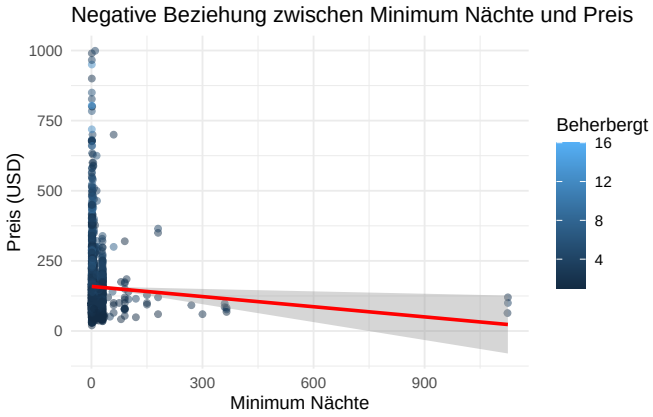


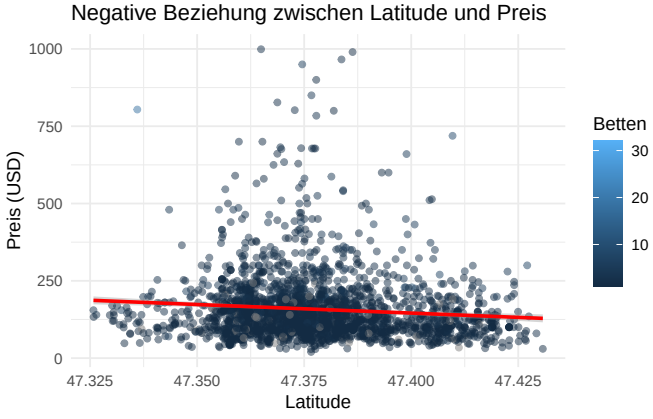


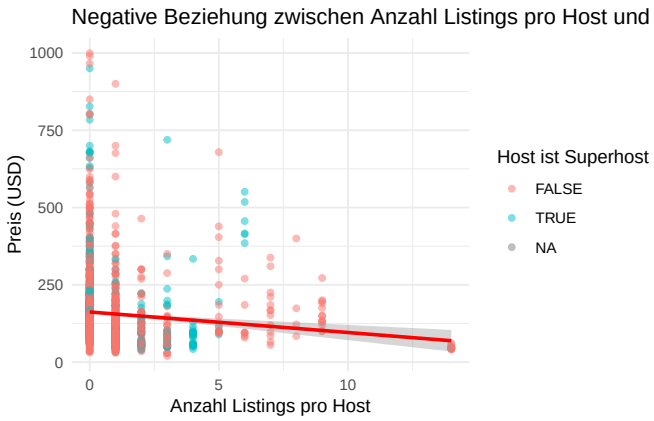
Die Analyse zeigt auf die irgendwie logische Korrelation zwischen den Anzahl Gästen respektive Betten und den Preisen eines AirBNBs... Also je grösser das AirBNB desto mehr kostet es...

Gibt es noch andere Korrelationen welche nicht so klar ersichtlich sind?

Vielleicht sind die negativen Beziehungen spannender:







Sieht schon ein wenig besser aus - vorallem die Latitude. Wobei auch hier ist es irgendwie klar. Je weiter weg das AirBNB von zentrum weg ist desto günstiger ist die Wohnung...

Gibt es sonst noch irgendwelche Möglichkeiten den Preis vorherauszusagen?

Descriptive analysis [7pt]

1. **Beschreibend (Descriptive)**: Deine Forschungsarbeit scheint beschreibende Analysen abzudecken, indem sie Merkmale von Airbnb-Unterkünften untersucht, die einen höheren Preis erzielen. Dies umfasst die Analyse der geografischen Lage, der Art der Unterkunft, der Anzahl der Zimmer und Bäder sowie der Verfügbarkeit von Annehmlichkeiten.

- Was reasonable statistics used for data inspection and which attributes have been checked? [2pt]
- Are the findings and statements qualitative? [3pt]
- What are the conclusions of the findings? [2pt]

Predictive analysis [7pt]

1. **Prädiktiv (Predictive)**: Um prädiktive Analysen einzubeziehen, könnte deine Fragestellung erweitert werden, um Vorhersagemodelle zu entwickeln, die den erzielbaren Preis basierend auf bestimmten Merkmalen der Unterkünfte prognostizieren. Dies könnte beispielsweise die Frage beinhalten: "Welche Merkmale einer Airbnb-Unterkunft sind prädiktive Indikatoren für einen höheren Preis?"

- Are the applied methods goal orientated? [2pt]
- Were the applied methods qualitatively performed? [3pt]
- What are the conclusions of the findings? [2pt]

Prescriptive analysis [7pt]

1. **Präskriptiv (Prescriptive)**: Um präskriptive Analysen abzudecken, könnte die Forschung Empfehlungen entwickeln, wie Gastgeber ihre Unterkünfte modifizieren könnten, um die Attraktivität und den Preis zu maximieren. Die Frage könnte lauten: "Welche Änderungen könnten Gastgeber vornehmen, um den Preis ihrer Airbnb-Unterkünfte zu maximieren?"

- Is the applied method reasonable for the given problem? [2pt]
- Were the applied methods qualitatively performed? [3pt]
- What are the conclusions of the findings? [2pt]

VI. 6. Ergebnisse (statistische Ergebnisse, Zahlen, Diagramme)

Conclusion [3pt]

- Are the initial BI problems reasonably discussed/solved? [2pt]
- Are the obtained new insights of the data sets of good quality? [1pt]

VII. 7. Schlussfolgerung (Beantwortung der Frage)

VIII. 8. Referenzen

```
[1] 2
```

```
[1] 4
```

The `echo: false` option disables the printing of code (only output is displayed).