BigData Missing Values Visualization Challenge

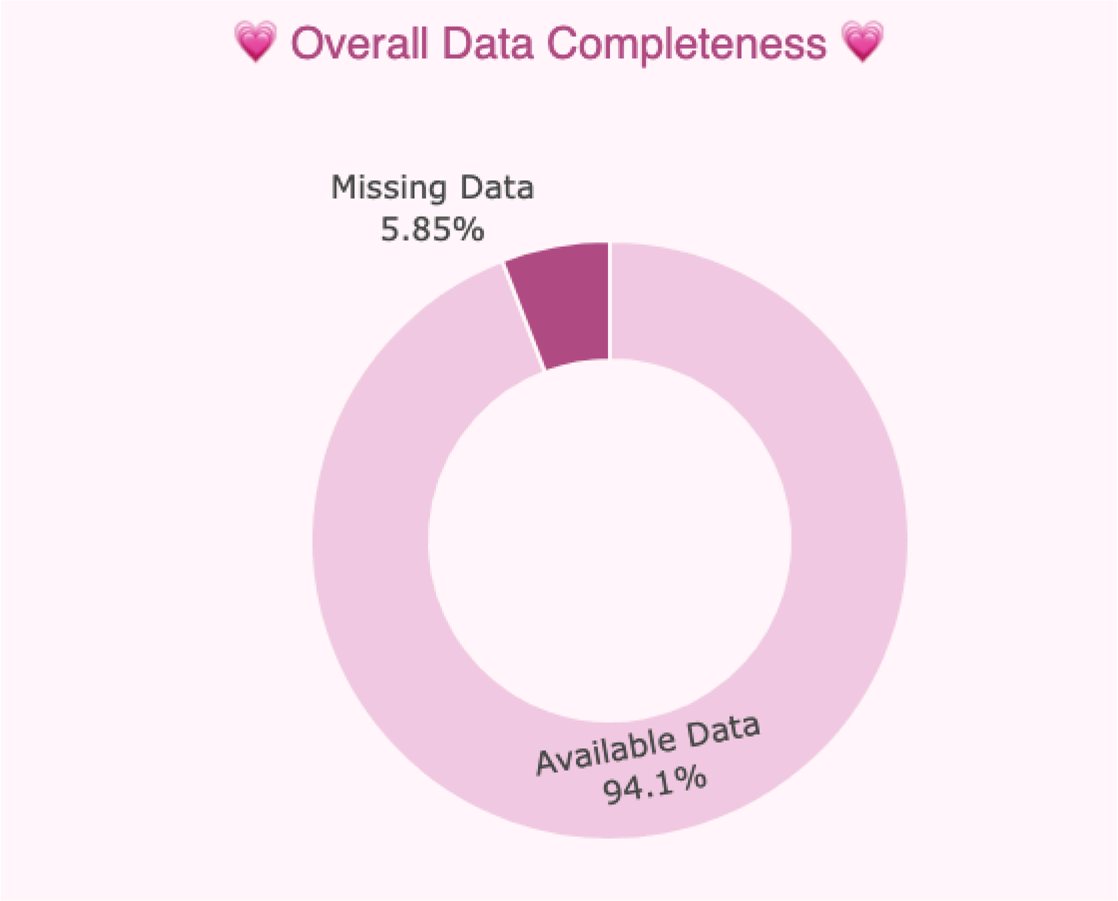
Team: Bohdan Hashchuk , Shvahuliak Kateryna , Hnyp Olena , Martsinkovska Anastasia, Strus Dmytro

In our implementation, these visualizations helps to discover data from different sides.

We displayed data from several different perspectives:

- Overall Data completeness
- Missing values by Feature
- Correlation of missing values between columns
- Trend of missing Ratings over time
- Missing ‘reviews_per_month’ by Superhost status
- Average Reviews per month by Bedrooms

Overall Data completeness



This visualization provides a **high-level overview of missing data** across the entire dataset. By summarizing completeness in a single, intuitive chart, it helps users quickly assess the overall data quality before diving into more granular analyses.

A compact percentage breakdown (available vs. missing) allows identifying whether data quality issues are minor or severe enough to require preprocessing or data collection improvements.

Key ideas:

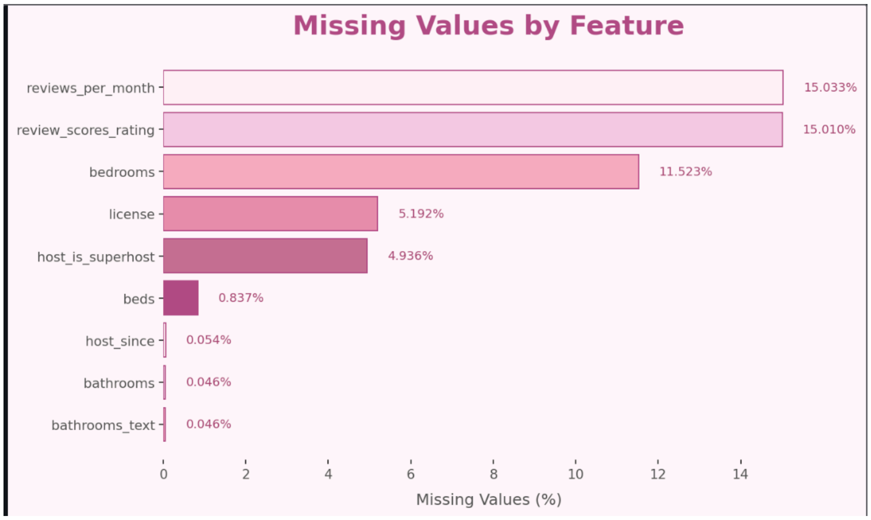
This considered to be a starting points for any data report, so we should not expect too much from it

Limitations:

- This visualization is not specific enough, we cannot see columns that have missing values

- Also, this assumes that all features are weighted equally, which is not true and may affect on precision of visualization
- This works perfect form small-medium datasets. bigger datasets may require more sampling

Missing values by Feature



While the overall completeness chart shows the *big picture*, this visualization helps **pinpoint which specific features contain missing data** and to what extent. It enables prioritization: features with higher percentages of missing values may require data imputation, removal, or additional collection, whereas those with negligible gaps can often be ignored during cleaning.

Key Ideas:

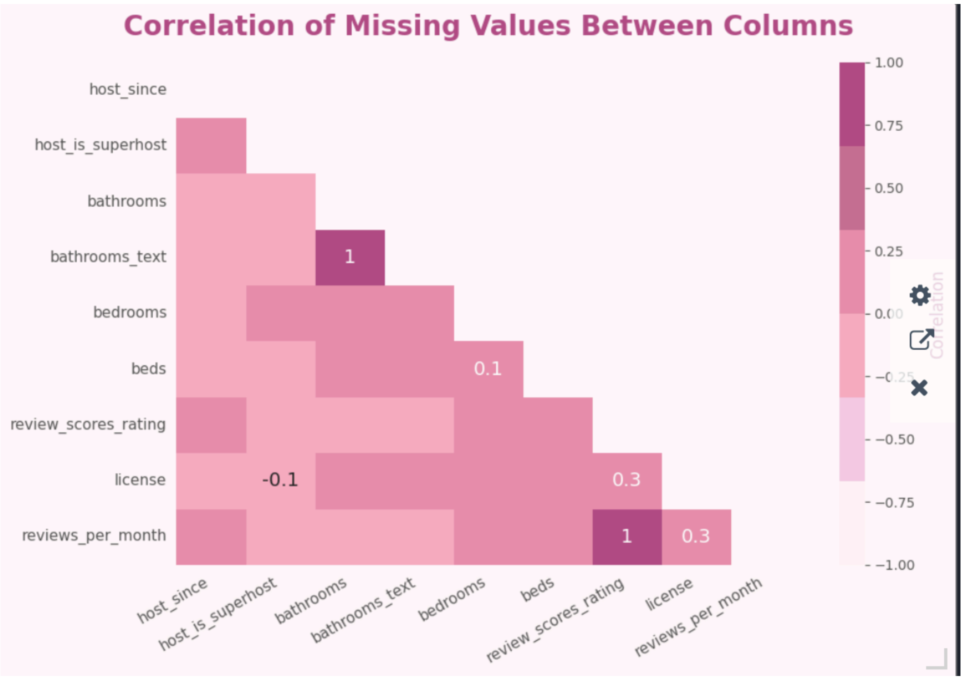
- Each bar represents a feature, the bar length corresponding to the percentage of missing data.
- We sorted features in descending order to see the most problematic ones
- This chart will help to clean the data more effficiently showing where analysts should focus on

Limitations:

- All the features are equally weightened again

- There is no info about inter-feature relationships to detect any possible dependecies.

Correlation of missing values between columns



After identifying which features have missing values, it's equally important to understand **whether those missing values occur together**. This visualization helps detect **patterns of dependency or shared causes** of missingness — for example, if two variables often have missing values in the same rows, it may indicate a common data entry issue or feature relationship.

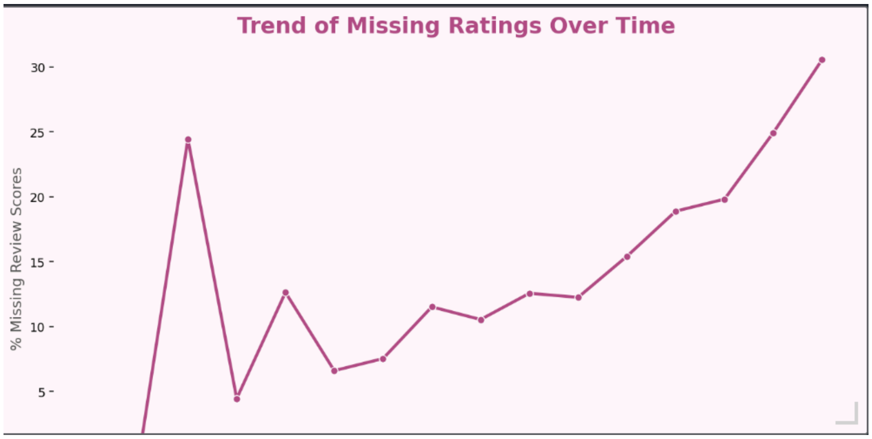
Key Ideas:

- Strong positive correlations suggest that when one column has missing data, the other likely does too.
- Negative correlations indicate inverse relationships — one column's missingness tends to correspond to another's completeness.

Limitations:

- When only a few values are missing, correlations can be unstable or misleading.
- Sometimes hard to understand

Trend of missing ratings over time



We have decided to understand whether missing ratings differ over time. And for our surprising, Yes! It differs. On the visualization me may clearly see that global trend is rising over time. So, the older item is ,the more likely it will have missing rating value

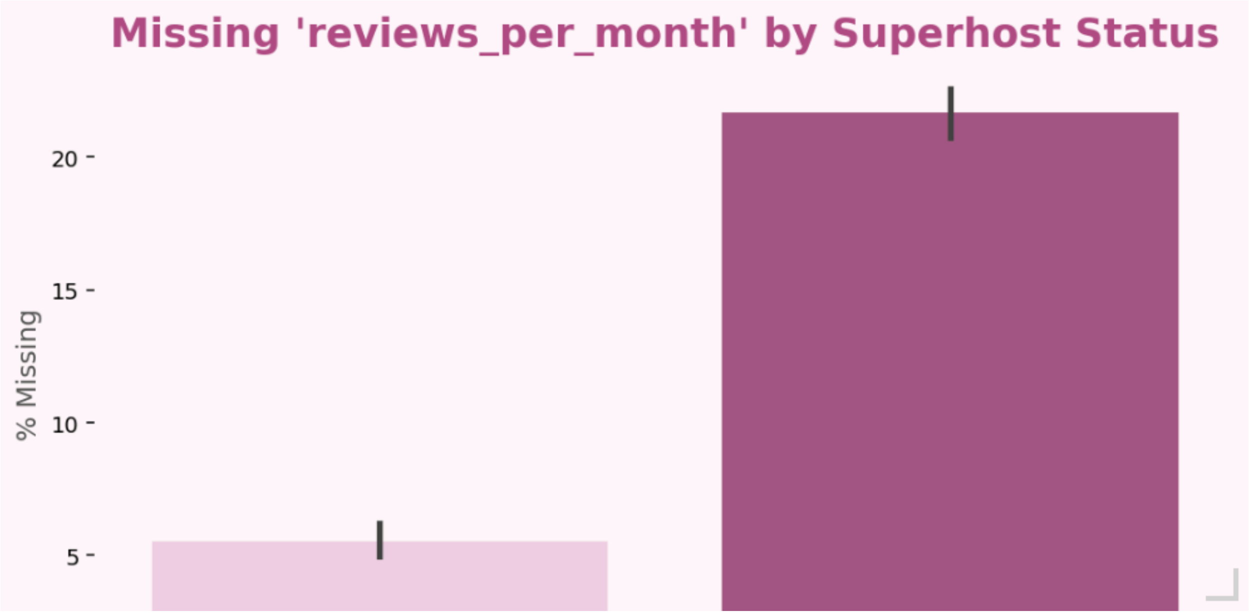
Key Ideas:

- Older listings demonstrate higher rates of missing ratings compared to newer ones
- The trend line makes it easy to identify critical time periods where data quality changed significantly

Limitations:

- The visualization doesn't explain *why* older listings have more missing ratings
- Temporal patterns may be confounded with other factors like seasonal variations or platform changes

Missing ‘reviews_per_month’ by Superhost Status



AirBnB has two different statuses of users. Host and SuperHost. SuperHost is verified and reliable host that has reviews, has been verified and many users prefer to choose Superhost over ordinary one. Thats why we decided to see whether missing values differ depending on this status. And yes, If host is Superhost it has 4 times less missing values rather than not verified host.

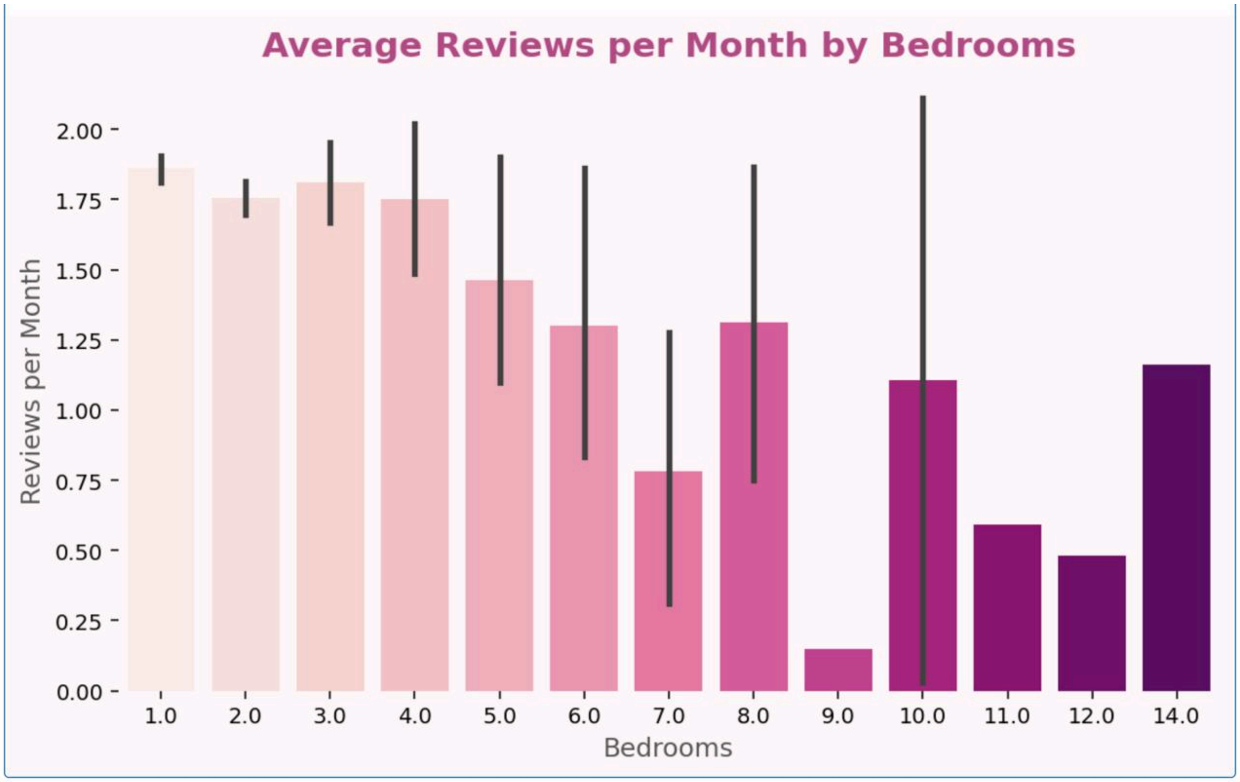
Key Ideas:

- Superhosts have approximately 4 times fewer missing values in 'reviews_per_month' compared to regular hosts
- This stark difference suggests that verified, engaged users maintain more complete profiles and receive more consistent reviews

Limitations:

- May oversimplify the relationship by not considering other factors like listing age, location, or property type that could influence both Superhost status and review frequency

Average Reviews per Month by Bedrooms



After our research with Superhost status, we decidedto go deeper and see the distribution of reviews depending on the amount of Bedrooms in the apartment. And we may see that the more bedrooms apartment has, less reviews it has. We can assume that up to 4 bedrooms apartments are more affordable so this pattern is pretty undertandable. But on the visualization may see that the more bedrooms apartment has , the bigger spread it is, so these values may significantly differ from month to month.

Key Ideas:

- Apartments with 1-4 bedrooms show higher review frequency, likely due to greater affordability and higher booking rates
- Variance increases dramatically for properties with more bedrooms, indicating inconsistent booking patterns for luxury or large properties and it makes it hard to predict

Limitations:

- High variance in larger bedroom categories makes it difficult to draw precise conclusions
- Doesn't account for confounding variables like location, price, or property type
- The visualization doesn't distinguish between missing reviews vs no bookings vs missing data entry
- Outliers in higher bedroom counts may be driven by a small number of properties