OKKA Health Analytics Engineer Assessment

1.

a. To start the data retention requirements should be specified, with regards to which historical data needs to be retained. Given the significant size of Airbnb's customer base and data volume full snapshots of data may need to be limited to specific use cases to not hinder processing performance.

   As such for the listings data it may be useful to use an incremental dbt materialisation for the staging source layer such that each month only new listings are ingested each month, and where any changes to a specific listing id can be updated.

   Using snapshots will allow changes to source data to be tracked within the Data Warehouse, this can provide valid_to and valid_from dates from which historical changes in listing parameters or host information can be stored. If there is not a updated_at field in the source table we can use dbt's check strategy over a specified set of parameters to check whether any of them have changed.

   Snapshots will also ensure that while historical changes in data are captured excessively large amounts of historical data are not retained hindering pipeline and processing performance.

b. Firstly when models/pipelines are created general best practices should be followed, for example avoiding using SELECT * from source tables if all the fields are not required.

   It is useful to add in the model configuration the on_schema *change* variable for incremental model to help aid schema change handling. Again, if the new columns are not required, this can be set to 'ignore', alternatively you can sync_all_columns to allow for all changes, append_new_columns to add new columns or fail in case further inspection is required.

c. Implementing dbt tests such to validate data in basic ways (unique, not null and accepted_values) is the first step to ensure that core data checks are made. Should this happen strategies should be drawn up to either provide default values or filter out these records.

   Given the connections between models it is also to inspect the dbt lineage graph to assess whether downstream models will be affected by such null or non-unique values to ensure future analyses are not misleading.

   Communications should also be made between other teams to assess the root cause of such data issues e.g. should a value be a required field on the front-end?

3. 9659
5. 215.0235602094241 characters
6. c)reviews with both hair dryers and washers
7.
To start with it is important to discuss with the management team what they see as the KPIs to monitor weekly performance for these properties, this will allow the agency to assess which data metrics are key to focus on for these dashboards.

Furthermore, we should contact the management team to determine their benchmarks to measure KPIs against. This will allow the data team to assess data sources for these dashboards. Perhaps more specifically any external data sources the team may need to build pipelines for to assess the KPIs against. As there may be more mature internal data pipelines these may require more time to develop.

An assessment of the internal data systems will also help inform the potential visualisation tools and warehouses in play.

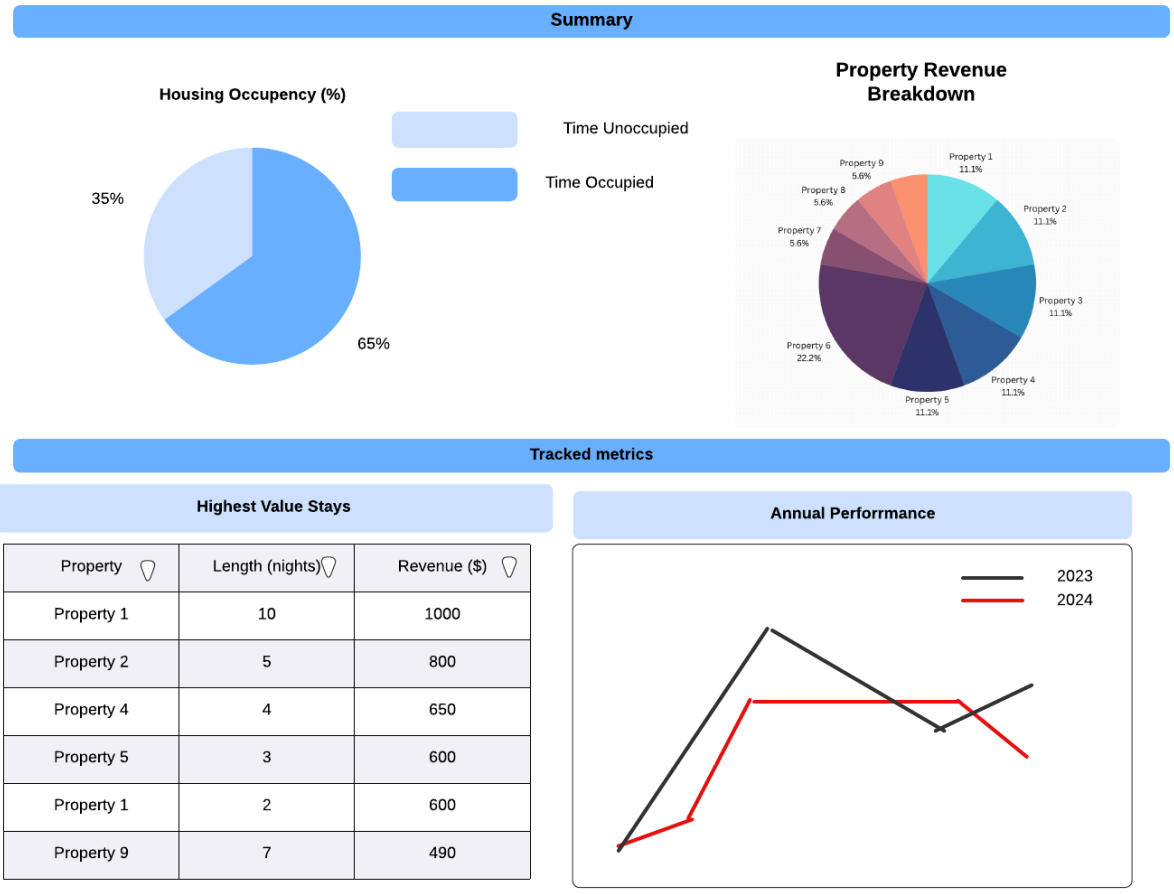
Moreover, while we know what cadence we will be assessing performance at (weekly) it is also important to understand what their preferred insights would be useful to the management team. If they want to see time based performance parameters the cadence over which they want to see that data is key too. This will inform the data team as to how to construct views over which the dashboards can be built, as well as which parameters can be used in incremental models to track historical changes.

From this, an amount of EDA should be performed to assess the data quality this will allow the data team to provide an early assessment of how much data cleaning will be needed and ask the management teams on potential strategies they may prefer to handle certain cases (null/accepted values).

From this further discussions should be made with the users to confirm their functional needs e.g. what do they want to be able to filter/sort by?

Following this the internal data team should have prioritisation discussions and scoping of the prerequisite tasks. This will allow the team to provide time estimates to delivery and key stages to assess the dashboard and take on feedback.

b.



C. External data of over revenue from AirBnb's around Amsterdam, narrowing down to revenue from similar sized properties etc. This could be used to compare how the overall financial performance is, as well as potential price benchmarking.

You could also look at features or amenities of listings in Amsterdam to see what features the agency's properties are missing compared to other successful properties around Amsterdam.

External data of transit rates from flights or ferries into Amsterdam could also be used to predict or forecast seasonal times when surge-pricing could be used.

d) As an example for the surge-pricing, comparative data of normalised price rises from transit into Amsterdam could be shown in comparison to seasonal price rises across the agency's properties. It should then be stressed the importance of correctly time-stamping the internal data.
As the agency knows what features each property has, aggregation could be made through external data to find that information from external Airbnb sources. One could also compare

external data’s postcode or co-ordinates to that of the agency’s properties to look at price comparisons compared to the proximity to more central areas of Amsterdam.