Assessment Exercise - BI Data Visualization Engineer

To the candidate:

Using this dataset which the candidate can download from our Economic Research site
URL: https://www.zillow.com/research/data/
Data Set:

- For-Sale Listings > For-Sale Inventory (Smooth, All Homes, Weekly)

Geography:

- Metro/US

1) Create a custom dashboard with multiple data visualizations using javascript or a language of your choosing. Do not use Tableau, Looker or similar tool offering built-in dashboarding capabilities. This dashboard should allow users to filter by state and date range, and answer the following questions:
    a) How is total for-sale inventory trending over time?
    b) What are the cities and states with the most available inventory?
    c) Assume you are a product manager in charge of tracking for-sale inventory across the US. What questions would you have and what visualizations would use to answer these questions?

2) Send the completed dashboard

3) Explain what design decisions you made and why (why did you choose certain visuals, what story does your dashboard tell)?

4) If you could have additional data:
    a) What other metrics would be useful here and why?
    b) How would you design a dataset to incorporate that data with the dataset above?

Answers

1c) Assume you are a product manager in charge of tracking for-sale inventory across the US. What questions would you have and what visualizations would use to answer these questions?

Some of the important questions and its corresponding visualizations would be:
1. What are the best regions in each state. For this I could use a visualization that offers a map and when the mouse hovers over a state the tooltip shows a small graph of inventory over time with the top 3 regions of the selected state.

2. A comparison to the week previous to the one being selected. Actually I added this one to the visualization, the best way to do it is to make a custom tooltip that indicates us which state are we selecting, the date and the change in %.
3. I would question what % of the total USA inventory represents each state. I would use this to see if there are states where the plataform is widely popular. A stacked bar graph where each bar represents a date can be useful.
4. I think having also which states/regions grew / lost the most comparing to last date is useful as we can know where there will be an increase in market in advance. A simple line graph with a % increase / decrease of house inventory (with red on the decreases and green on increases) can be useful.

Explain what design decisions you made and why (why did you choose certain visuals, what story does your dashboard tell)?

The first visuals I decided were the cards, there is no better way to show the "top" of something than simple information, but I needed it to change whenever the dates are changed, so for it to not be confusing I added the date information in the cards to know which were the best regions and states.

The line graph is always a good option when talking about time series. I needed the tables to be dynamic to show all the dates in the date ranges. Part of the things that were needed to be added was a tooltip that tells us the actual numbers that were happening on each date, but numbers themselves are nothing if we do not have any comparisons. So, there are two things that were added. First, when you change the end date of the date ranges it will always select the top 5 states of said end date ordered from highest to lowest, with the possibility of adding any other state using a checkbox including the USA in order to know how it behaved in comparison to the country's performance. Second, I added a percentage in change from previous week in the tooltip to know what the change within the same state was.

If you could have additional data:

a) **What other metrics would be useful here and why?**
b) **How would you design a dataset to incorporate that data with the dataset above?**

A) Other good metrics would be to have been to know how many houses are in construction on each week, this to know which the construction to for-sale convert rate is within each region to see if the market is shift more to for-rent instead of for-sale.

Other information that would be useful is the reasons of a for-sale house to be out of inventory, this is mainly to know if there are negative reasons for a house leave the inventory (for example a positive one is that it is out of the catalogue because a hose was sold) in order to know if we are performing good or bad.

Other one would be a dataset that contains a forecast of the next week's inventory using statistics or AI models; this would be useful for having an action plan for the foreseeable future and when the dashboard updates with next week we can have a measurement of how accurate was our forecast.

Other one would be a metric that indicates the population of each state, this can be useful to measure the proportion of citizens vs available houses for sale to see if there can be problems with an excess of houses or see if people are not buying houses.

Other one would be a tag indicating the GDP per capita of each region along with the median price of each region, this mainly to see if the prices of the houses make sense with the income of the people living in a certain region.

B) For the first one I would make a file exactly as the extracted one but with the correct numeric information and changing headers; having the information in same structure and granularity is always useful when we are comparing time-based metrics on the same base.

For the other one I would make a table structured like house_id - unsubsribed_date (preferably the corresponding Saturday to match it with the extracted data) – unsubscribe_reason (limited response options) – classification (good or bad according to the reason).

The forecast data would be to add a register to the extracted file but with a tag that indicates that it is a prediction for us to not take it into account as a real sales number.

For the population in each state, we would structure it as follows census_year-state in another table, it would be this way because the updates in this data is not measured constantly so we have to use what we get. We can integrate it by just adding a current_population header in the same table.

For the GDP information, the best way is to have another table to join it that has the region_id-measurement_date-last_measured_gdp-house_median_price, and then join this with the original table.