## Statistical/Hypothetical Question

### Outcome of EDA

Based on the results of Exploratory Data Analysis (EDA), it is evident that there is a strong correlation between the square footage of living space (area) and the Rent of the house(rent).

Including additional variables such as the number of bathrooms, bedrooms, whether rental house present inside a gated community, availability of amenities such as parking, power backup to the analysis, we observe a vigorous positive relationship with the rent variable.

Correlation Coefficient of number of bedrooms vs rent is 0.69. This indicates that a strong positive correlation.

Regression Analysis using multiple explanatory variables revealed that 54% variability in Rent (R-squared value: 0.54) can be explained by variables such as number of bedrooms, bathrooms, availability of parking, furniture, and power backup. Skew value of 1.923 indicates a right-skewed distribution.

***What do you feel was missed during the analysis?***

Though EDA helped highlighting strong correlation between certain variables, null hypothesis being rejected, skewness of distribution, the study was performed on selected variables. Analysis of all possible variables including the ones that were not selected for the study would have revealed more surprising findings. Hypothesis Test resulting in “non-rejection” of hypothesis could have been more satisfying.

***Were there any variables you felt could have helped in the analysis?***

Availability of other critical amenities such as elevator, air conditioning could have thrown more light into the factors determining rent. More experience in performing EDA on robust and dynamic datasets could have added more confidence to result of my studies.

***Were there any assumptions made you felt were incorrect?***

For the scope of the study, I have assumed the area variable to denote the total square feet of Living area.

Analysis of just ~5300 rental records to understand the factors of rent on a country with 1.5 Billion population (with estimated 200 million people living in metro cities) might make the results of the study, a significant assumption.

Selection of variables for the study have been made based on the assumption that they are strongly correlated variable. This assumption can be verified by performing study on remaining variables.

***What challenges did you face, what did you not fully understand?***

Finding the suitable dataset with required number of quantitative variables, at least one categorical variable, and at least sizable number of records was challenging. Finding the ideal variables for various exercise such as PMF, CDF, Hypothesis Test, etc., was challenging, I have adopted a trial and error method before settling on the variables for some of the exercises.