### 1. How did you determine the size of the different organizations?

Answer:  
 We did not have a direct variable for organizational size. However, we inferred size based on agency groupings provided in the dataset under the agency variable. We categorized them as large vs. small agencies using pre-known agency scales (e.g., DoD and VA considered large; smaller bureaus categorized accordingly). For improved clarity, this assumption will be stated explicitly.

### 2. How could you say that "Older employees (DAGEGRP = B) and supervisors (DSUPER = B) showed higher satisfaction"?

Answer:  
 This statement is based on group-wise comparison using boxplots and group means. In the EDA phase (CHA\_1\_EDA.ipynb), we segmented job satisfaction scores by DAGEGRP and DSUPER. Employees in group B consistently showed higher mean satisfaction.

### 3. Is there any result supporting this explanation?

Answer:  
Yes. We visualized these insights through group-level plots (boxplots and bar plots) and also included group-wise average comparisons in the EDA. The difference was not only visual but also supported by statistically significant mean differences (t-test and ANOVA in exploratory stage).

### 4. How did you decide what the most important variables are?

Answer:  
 Variable importance was determined by:

* Standardized regression coefficients (β) in linear models
* Feature importance scores from ML models (Random Forest and Gradient Boosting)
* Additionally, SHAP summary plots were generated (in CHA\_2\_Normal.ipynb) to interpret non-linear model contributions

5. R² is 0.48 — are there other variables that you did not consider?

Answer:  
 Yes. While our R² = 0.48 indicates moderate explanatory power, other variables such as:

* Emotional exhaustion (burnout-related)
* Team dynamics beyond direct supervisor
* External work stressors  
   were not included due to either unavailability in the dataset or high missingness. This will be noted as a limitation in our conclusion.

### 6. Why did you not consider other factors that would have improved the model’s variance?

Answer:  
 We prioritized features that were consistently available and statistically clean (low missing data, interpretable). Including too many low-quality variables could reduce model validity. Future work will incorporate more complex psychosocial and behavioral variables as they become available in later survey versions.

7. Can you include regression and residual plots in Slide 16?

Answer:  
 Yes. We will update Slide 16 to include:

* Standardized residuals vs. fitted values plot (to test homoscedasticity)
* QQ plot (to test normality of residuals)
* Regression line with confidence interval

These are already present in CHA\_1\_Normal.ipynb and will be visualized appropriately.

### 8. How can you say that larger agencies exhibit stronger leadership-retention effects?

### Answer: This statement was based on stratified regression outputs where agency categories (large vs. small) were added as control variables. Larger agencies like DHS and DoD showed stronger β coefficients between Leaders Lead and intent to stay (DLEAVING). We'll make sure this is backed by explicit numbers in our final submission.

9. Need proper explanation for findings in page 20

Answer:  
 Page 20 discusses POCI regression results. We’ll revise this to explain that:

* Q15 (coworker cooperation) was the strongest predictor (β = 0.1837)
* Other important variables relate to disruption handling, innovation, and team climate
* DRNO (Race/National Origin) showed variation in POCI strength, with R² ranging from 0.5184 to 0.5815  
   A full sentence explanation will be added for each metric.

10. Why did you put same graphs in slides 17 & 21?

Answer:  
 That was an error in slide duplication. One of them was intended to show feature importance (Slide 17) and the other the SHAP plot (Slide 21), but we mistakenly reused the same Random Forest bar chart. We will replace Slide 21 with a SHAP or Gradient Boosting plot and add distinct interpretation.