


The background of the slide is a light gray gradient. It is decorated with numerous realistic water droplets of various sizes. Some droplets are at the top left, some are scattered in the middle, and a larger cluster of droplets is on the right side. The droplets have highlights and shadows, giving them a three-dimensional appearance.

PIZZAZ.COM DATA ANALYSIS

PREPARED BY JOANNE MUSA



OUTCOMES

- DETERMINE BEST PREDICTORS FOR “LIKE” TO OPTIMALLY MATCH COUPLES
 - DETERMINE A BEST EQUATION TO PREDICT “LIKE”
 - DETERMINE IF AGE DIFFERENCE IS SIGNIFICANT
 - DETERMINE IF RACE IS SIGNIFICANT
- 

DESCRIPTIVE STATISTICS

- MEAN AGE OF MALES AND FEMALES IS ABOUT 26 YEARS OLD
- THE MEAN AGE DIFFERENCE OF DATING COUPLES IS ABOUT 3.8 YEARS AND TYPICAL VARIATION OF ABOUT 3.2 YEARS.
- MALES AND FEMALES HAVE VERY SIMILAR MEAN SCORES FOR ALMOST ALL VARIABLES
- OF THE 63 SUCCESS DATES (BOTH MALE AND FEMALE MUTUALLY WANTED TO SEE EACH OTHER AGAIN), ATTRACTIVE, SINCERE, INTELLIGENT, AND FUN HAD THE HIGHEST MEAN SCORES
- THE HIGHEST CORRELATED VARIABLES TO LIKE ARE ATTRACTIVE AND FUN
- 119 DATES ARE OF THE SAME RACE, 151 DATES ARE OF DIFFERENT RACES, 6 DATES ARE INCONSISTENT WITH SAME RACE DUE TO MISSING VALUES

BUILDING THE BEST MODEL

- MALE AND FEMALE DATA ARE COMBINED FOR LIKE, ATTRACTIVE, SINCERE, INTELLIGENT, FUN, AMBITIOUS, AND SHAREDINTERESTS TO CREATE A SINGLE BEST MODEL
- TWO ADDITIONAL VARIABLES ARE CREATED: AGE_DIFF & SAMERACE
- POLYNOMIAL TERMS ARE CONSIDERED TO ACCOUNT FOR ANY CURVATURE TO THE DATA
- INTERACTION TERMS ARE CONSIDERED TO DETERMINE ANY RELATIONSHIPS THAT MAY EXIST BETWEEN THE PREDICTOR VARIABLES
- FOUR METHODS FOR FITTING THE BEST MODEL IS USED: THE STEPWISE FORWARD APPROACH FITTED THE BEST MODEL
- R^2 OF THE BEST MODEL = 0.67, CONTAINS 4 VARIABLES AND 5 TERMS

$$\hat{y} = 5.69536 + 0.02080X_{18} + 0.01401X_{23} - 0.18506X_{40} + 0.24564X_{42}$$


where \hat{y} = Predicted Like value, X_{18} = Attractive*Fun, X_{23} = Sincere*Intelligent,
 X_{40} = Ambitious*SameRace, X_{42} = SharedInterests*SameRace

DIAGNOSTICS

- DATA ERRORS: NO DATA ENTRY ERRORS APPARENT, SAMPLE SIZE = 276, MISSING DATA ARE LISTED (TABLE 3)
- OUTLIERS: JACKKNIFE RESIDUALS & BOXPLOT METHODS EMPLOYED
 - BOXPLOT: OBSERVATIONS **116, 151, 252, 260** ARE IDENTIFIED AS OUTLIERS
 - JACKKNIFE RESIDUALS: OBSERVATIONS 50, 62, 97, **116**, 133, **151, 252, 260** ARE IDENTIFIED AS OUTLIERS
 - HIGH OUTLIERS: OBSERVATIONS 62, 97, AND 133 ARE HIGH OUTLIERS IDENTIFIED BY JACKKNIFE RESIDUALS ONLY; INVESTIGATION OF THESE OBSERVATIONS MAY BE CONSIDERED DUE TO CONCERNS OF INFLATED SCORES
- ASSUMPTIONS: NORMALITY, $E(e_o) = 0$, HOMOGENEITY, AND INDEPENDENCE ARE ALL SATISFIED
- COLLINEARITY: NO COLLINEARITY PROBLEMS WITH THE BEST MODEL



RELIABILITY

- SPLIT SAMPLE ANALYSIS (CROSS VALIDATION) IS USED TO ASSESS RELIABILITY OF THE MODEL.
 - RESULTS OF THIS PROCESS CONFIRMED THAT THE BEST MODEL IS RELIABLE
- 

CONCLUSION

- THE BEST MODEL IS $\hat{y} = 5.69536 + 0.02080X_{18} + 0.01401X_{23} - 0.18506X_{40} + 0.24564X_{42}$

*where \hat{y} = Predicted Like value, X_{18} = Attractive*Fun, X_{23} = Sincere*Intelligent,
 X_{40} = Ambitious*SameRace, X_{42} = SharedInterests*SameRace*

- AGE DIFFERENCE IS NOT SIGNIFICANT
- OPTIMIZING MATCHES FOR FUN, ATTRACTIVE, SINCERE, AND INTELLIGENT WILL INCREASED PREDICTED LIKE
- SAME RACE IS SIGNIFICANT, BUT NOT INDEPENDENTLY: SAMERACE INTERACTS WITH AMBITIOUS AND SHAREDINTERESTS
- OPTIMIZING SAME RACE MATCHES IN CONJUNCTION WITH OPTIMIZING SHAREDINTERESTS WILL INCREASE PREDICTED LIKE