Skewness:
measure of the asymmetry of the probability distribution – source of bias pulls results in one direction; discuss skewness statistic; acceptable range +/-2(1.96); +/-3.29 depending on advice. Should recognise this as an issue for majority of variables.

Kurtosis:
measure of whether the data are peaked or flat relative to a normal distribution – source of bias; should discuss kurtosis statistic; acceptable range +/-2 – should identify kurtosis of length of stay

Prediction:
looking at is the variance in the outcome variable and how much of the variance could be considered to be explained by the predictor variables – should be applied to problem

Coefficients:
Measure the strength and direction of relationships between independent variables and the dependent variance – need to consider the significance of these

Representativeness:
sample is drawn from population; size and randomness are issues; needs to reflect the population in terms of key characteristics and variables being investigated; source of bias and missing data. Should consider the sample frame as described – issues exist.

Outliers:
Cases with values very different to majority of data; a lot of statistical tests are sensitive to outliers; source of bias; can affect both assumptions of tests and results.
Expect discussion of box plots and zscore frequency tables could also consider histograms; 95% to fall within +/-1.96 none greater than +/-3.29. Outliers do exist and are an issue

Outliers:
Multicollinearity can be greatly influenced by outliers - They ‘pull’ the equation away from the general pattern Sometimes it is reasonable to delete the outliers; Sometimes you need to retain the outlier – it may be your most important observation; Sometimes it is a good idea to do your analysis twice – once with and once without the outlier to see how much influence it is having on the model
suitability of the dataset

Sample size, sample to variable: generally at least 300 cases needed; N:p ratio where N = no of cases, p=no of variables rule of thumb 10:1. Not an issue here. Alternate rules may be accepted; There are some concerns here but 100+ is acceptable

Central Tendency and Variability:

single value that attempts to describe a set of data by identifying the central position within that set of data. Expect brief discussion of Mean, Median, Mode

most representative value and spread of data around this; important for description but also statistical tests depend on these two concepts;

Dispersion:

how the data is dispersed around the central measure; Expect brief discussion of standard deviation, interquartile range, range.

Type I error:

incorrect rejection of null hypothesis – false positive;
Should relate to statistical significance - alpha is the maximum probability that we have a type I error. For a 95% confidence level, the value of alpha is 0.05. This means that there is a 5% probability that we will reject a true null hypothesis. Should link to skew.

Type II error:

failure to reject null hypothesis – false negative.
Should relate to statistical power: probability of correctly rejecting a false null hypothesis; Should consider sample size.

Hypothesis testing steps:

Derive null and alternate hypotheses;
identify type of test needed based on hypothesis;
decide on cut-off and critical value;
conduct the test to compute test statistic;
interpret statistical significance and test statistic;
conclude on whether to accept/reject null hypothesis.

Coefficients:

Measure the strength and direction of relationships between independent variables and the dependent variance – need to consider the significance of these. Both significant – should explain unit of change.

F statistic:

Whether the model as a whole predicts the dependent variable; Its statistical significance is the significance of the model; Model is significant

interaction effect

An interaction occurs when the magnitude of the effect of one independent variable (X) on a dependent variable (Y) varies as a function of a second independent variable (Z).

Regression

will allow researcher to explore how well a set of variables is able to predict an outcome variable; which variable in a set is the best predictor; whether a variable is still able to predict an outcome when controlling for a particular variables;

Normality of residuals:

Should follow a normal distribution with a mean of 0; Since the residuals measure where the points fall in relation to the regression line, a normal distribution of residuals indicates that the same number of points fall above and below the line; Look at histogram and residual statistics. (Appears to be a problem – need to consider outliers.)

Prediction:

looking at is the variance in the outcome variable and how much of the variance could be considered to be explained by the predictor variables. Should comment on variables involved

Dummy variable:

interested in effect of characteristic, if include 1,2,3,4,5 etc. will act as multiplier, 0 and 1 can see effect and its absence; should address need to recode into multiple variables

Multicollinearity:

Occurs when two or more independent variables contain strongly redundant information; If variables are collinear then it means there is not enough distinct information in these variables for MLR to operate – they are essentially measuring the same thing; Examining a correlation matrix that compares your independent variables with each other; coefficient above 0.8 suggests collinearity might be present

Missing Data:
Need to identify reason for missing; Need to care because source of bias, could reduce statistical significance and power, increase risk of type I and type II; Should address MCAR, MAR, MNAR, distinguish between then and discuss detection: MCAR Little's test, patterns to infer MAR, if cannot find reason from pattern then MNAR; Options – delete, impute, repeat analysis with and without the data; Expect discussion of info provided: numbers missing, Little's test.

Principal Components Analysis

Objectives: Reduce the number of variables; Evaluates the construct validity of a scale, test, or instrument; Development of parsimonious (simple) analysis and interpretation; Addresses multicollinearity (two or more variables that are correlated)
Used to reduce multiple observed variables into fewer components that summarize their variance; All of the variance in the variables is used. Reveals principal components underlying structure of the data

Extraction: aim is reduce a large number of items into factors. Need to decide how many factors to retain. Criteria for extraction: Kaiser's criteria (eigenvalue > 1 rule), the Scree test, the cumulative percent of variance extracted. Rotation: Unrotated results from a factor analysis are not easy to interpret, although the plot helps. Rotation helps clarify and simplify the results of a factor analysis. Conceptually, the axes are being rotated so that the clusters of items fall as closely as possible to them. Making the factor loading pattern much clearer as one of the two pairs of coordinates for each item tends to be close to 0.00

Variance: Looking at total variance explained for factors with eigenvalues > 1; component 1 explains 46.89% of total variance; single component Communalities: proportion of the common variance within a variable; looking at extraction column – all above .3 Except for good standard of care; Should look at component loadings before rotation and after to see difference Should draw conclusions about the number of components and variance explained - single component – conforms to what researcher was expecting.

Component: uncorrelated variable constructed from a set of correlated variables by considering pattern of co- variance – represent underlying structures within data contained in set of correlated variables. principal components are the eigenvectors of the covariance matrix of the original dataset(Here there is one – self-esteem. Cronbach's alpha – index of

reliability; Reliability accuracy and precision of a measurement procedure; associated with the variation accounted for by the true score of the "underlying construct"; Acceptable level depends on the purpose of the instrument. Acceptable 0.60. For this data it is .92 – extremely strong
KMO: proportion of variance in variables that might be caused by an underlying factor. High values close to 1 suggest that FA useful, anything less than 0.5 FA won't be useful. This data value is 0.93 – Extremely good
Bartlett tests that variables are correlated. Looking for a statistically significant p value which suggests that they are related and a FA would be useful. No issues for this data.

Communalities: proportion of the common variance within a variable; Initial - By definition, the initial value of the communality in a principal components analysis is 1. Extraction - The values in this column indicate the proportion of each variable's variance that can be explained by the principal components. Variables with high values are well represented in the common factor space, while variables with low values are not well represented. Should draw conclusions about the number of factors and variance explained. All above .6 loading onto one component.

Eigenvalues: are the variances of the components. A larger eigenvalue means that that principal component explains a large amount of the variance in the data. A component with a very small eigenvalue does not do a good job of explaining the variance in the data. Total Variance: Looking at total variance explained for factors with eigenvalues > 1; 1st component accounts for most variance, next as much of that left over and so on; 1 component explaining 55.9% of total variance

why Principal Component Analysis is the best approach to use for data reduction in this inquiry rather than Exploratory Factor Analysis.

explanation of commonality – both reduce number of variables
Both aim to achieve data reduction and help in development of parsimonious (simple) analysis and interpretation

If you were asked to choose between exploratory factor analysis and principal component

Objectives: Reduce the number of variables; Development of parsimonious (simple) analysis and interpretation;
Differences/Use EFA and PCA: EFA Determine the nature of and the number of latent variables that account for observed variation and covariation among set of observed indicators; Used to develop theoretical constructs;; PCA Reduce multiple observed variables into fewer components that summarize

their variance;
Should choose EFA : Used to prove/disprove proposed theories;

Exploratory Factor Analysis

Objectives: Reduce the number of variables; Evaluates the construct validity of a scale, test, or instrument;
Development of parsimonious (simple) analysis and interpretation; Addresses multicollinearity (two or more variables that are correlated)
EFA Determines the nature of and the number of latent variables that account for observed variation and covariation among set of observed indicators; Used to develop theoretical constructs; Used to prove/disprove proposed theories; only shared variance is used;
Sample size, sample to variable: generally at least 300 cases needed;
KMO proportion of variance in variables that might be caused by an underlying factor. High values close to 1 suggest that FA useful, anything less than 0.5 FA won't be useful. Bartlett tests that variables are correlated. Looking for a statistically significant p value which suggests that they are related and a FA would be useful. No issues for this data.
Reliability accuracy and precision of a measurement procedure; Cronbach's alpha – index of reliability associated with the variation accounted for by the true score of the "underlying construct"; Acceptable level depend on the purpose of the instrument. Acceptable 0.60. For this data it is 0.716. No significant improvement by deleting items.
Extraction: Need to decide how many factors to retain. Criteria for extraction: Kaiser's criteria (eigenvalue > 1 rule), the Scree test, the cumulative percent of variance extracted. 2 factors Communalities: proportion of the common variance within a variable; looking at extraction column Should look at factor loadings before rotation and after to see differences if any
Rotation: Unrotated results from a factor analysis are not easy to interpret, although the plot helps. Rotation helps clarify and simplify the results of a factor analysis. Conceptually, the axes are being rotated so that the clusters of items fall as closely as possible to them. Making the factor loading pattern much clearer as one of the two pairs of coordinates for each item tends to be close to 0.00

assumptions ***that must be met for the independent and dependent variables when undertaking a*** *linear regression.*

Variable Type: Outcome must be continuous; Predictors can be continuous or dichotomous. Non-Zero Variance: Predictors must not have zero variance. Linearity: The relationship we model is, in reality, linear. Should refer to tests in Q1. Independence: All values of the outcome should come from a different case.

interaction effect

An interaction occurs when the magnitude of the effect of one independent variable (X) on a dependent variable (Y) varies as a function of a second independent variable (Z).
Add an interaction term to the model – new variable is an interaction term, allows exploration of whether the effect of risk is different for different ages

differences between PCA and exploratory factor analysis.

Differences/Use EFA and PCA: EFA Determine the nature of and the number of latent variables that account for observed variation and covariation among set of observed indicators; Used to develop theoretical constructs; Used to prove/disprove proposed theories; only shared variance is used; PCA Reduce multiple observed variables into fewer components that summarize their variance; All of the variance in the variables is used

Cook's distance: Measures the effect of deleting a case; Look at Cook's distance for values greater than one – no major impact by deleting

Chi-square test

Working with nominal data; Can't use mean must use frequencies;
The null hypothesis is used to construct an idealized sample distribution of expected frequencies that describes how the sample would look if the data were in perfect agreement with the null hypothesis - that the proportions are the same for all populations
A chi-square statistic is computed to measure the amount of discrepancy between the ideal sample (expected frequencies from H0) and the actual sample data (the observed frequencies).