/\* Import the CSV file \*/

**proc** **import** datafile='C:\Users\jyo\desktop\TB\_burden\_age\_sex\_2023-10-28.csv'

out=TB\_burden

dbms=csv

replace;

getnames=yes;

**run**;

/\* Check for missing data \*/

**proc** **freq** data=TB\_burden;

tables \_all\_ / missing;

**run**;

**proc** **freq** data =TB\_burden\_clean;

tables age\_group\*risk\_factor / chisq cmh;

where age\_group in ('0-14', '15-24', '25-34', '35-44', '45-54', '55-64', '65+');

**run**;

/\* Impute missing data or remove missing data - Example of removing missing data \*/

**data** TB\_burden\_clean;

set TB\_burden;

if cmiss(of \_all\_) = **0**;

**run**;

/\* Descriptive Statistics and Visualization for Continuous Variable 'best' \*/

**proc** **means** data=TB\_burden\_clean n mean std min q1 median q3 max;

var best;

**run**;

/\* Bar Chart for Age Groups vs Risk Factors \*/

/\* This bar chart shows the distribution of different risk factors across age groups. \*/

**Proc sgplot** data=TB\_burden\_clean;

vbar age\_group / group=risk\_factor response=best stat=mean;

title "Distribution of Risk Factors Across Age Groups";

run;

/\* Heatmap for Age Groups and Risk Factors:

the heatmap can be useful to visualize the intensity of TB burden across different age groups and risk factors. \*/

**Proc sgplot** data=TB\_burden\_clean;

heatmap x=age\_group y=risk\_factor / colorresponse=best colormodel=(blue yellow red);

title "Heatmap of TB Burden by Age Group and Risk Factor";

run;

**proc univariate** data=TB\_burden\_clean;

var best;

histogram best / normalcurve;

inset n mean std / pos=ne;

run;

/\* Descriptive Statistics and Visualization for Categorical Variable 'age\_group' \*/

**proc** **freq** data=TB\_burden\_clean;

tables age\_group / nocum nopercent missing;

**run**;

/\* PROC SGPLOT procedure to create a vertical bar chart \*/

**proc** **sgplot** data=TB\_burden\_clean;

vbar age\_group / response=best stat=mean;

**run**;

/\* Descriptive Statistics for 'best' Grouped by Categorical Variable 'age\_group' \*/

**proc** **means** data=TB\_burden\_clean n mean std min q1 median q3 max;

class age\_group;

var best;

**run**;

**/\* Statistical graphics \*/  
proc** **sgplot** data=TB\_burden\_clean;

vbox best / category=age\_group;

**run**;

/\* Test of Normality for 'best' \*/

**proc** **univariate** data=TB\_burden\_clean;

var best;

histogram best / normalcurve;

qqplot best / normal(mu=est sigma=est);

**run**;

/\* One-Sample T-Test for 'best' \*/

**proc** **ttest** data=TB\_burden\_clean h0=**10000**;

var best;

**run**;

/\* Chi-Square Test for Independence with Odds Ratio and Relative Risk \*/

**proc** **freq** data=TB\_burden\_clean;

tables sex\*risk\_factor / chisq relrisk;

**run**;

/\* ANOVA for 'best' by 'sex' \*/

**proc** **glm** data=TB\_burden\_clean;

class sex;

model best = sex;

means sex / tukey; /\* Tukey's honestly significant difference test \*/

**run**;

/\* Scatter Plot and Correlation Test for 'best' and 'lo' \*/

**proc** **sgscatter** data=TB\_burden\_clean;

plot best\*lo;

**run**;

/\* correlation coefficients between variables in a dataset\*/

**proc** **corr** data=TB\_burden\_clean;

var best lo;

**run**;

/\* Simple Linear Regression with Diagnostics \*/

**proc** **reg** data=TB\_burden\_clean;

model best=lo;

plot r.\*p. rstudent.\*p. / cooksd;

**run**;

/\* Multiple Regression Analysis with Categorical Predictors \*/

**proc** **glmselect** data=TB\_burden\_clean plots=all;

class age\_group sex risk\_factor; /\* Declare categorical variables \*/

model best=lo hi age\_group sex risk\_factor / selection=stepwise; /\* Stepwise selection \*/

output out=reg\_output p=predicted r=residuals; /\* Output dataset with predicted values and residuals \*/

**run**;

/\* Plot Residuals \*/

**proc** **sgplot** data=reg\_output;

scatter x=predicted y=residuals;

refline **0** / axis=y;

**run**;

/\* Check the levels of the 'sex' variable \*/

**proc** **freq** data=TB\_burden\_clean;

tables sex;

**run**;

/\* Create a binary outcome variable for logistic regression \*/

**data** TB\_burden\_clean;

set TB\_burden\_clean;

if best > **10000** then outcome=**1**;

else outcome=**0**;

**run**;

/\* Logistic Regression Analysis with Model Fit Statistics \*/

**proc** **logistic** data=TB\_burden\_clean desc;

class sex(ref='a') age\_group(ref='0-14') risk\_factor(ref='smk') / param=ref;

model outcome(event='1') = lo hi age\_group sex risk\_factor / selection=stepwise lackfit;

output out=predicted\_values p=probabilities;

**run**;

/\* ROC Curve \*/

**proc** **logistic** data=TB\_burden\_clean plots(only)=roc;

class age\_group(ref='0-14') sex(ref='a') risk\_factor(ref='smk') / param=ref;

model outcome(event='1') = lo hi age\_group sex risk\_factor;

**run**;

/\* Independent t-test for 'best' between two age groups \*/

**proc** **ttest** data=TB\_burden\_clean;

class age\_group;

var best;

where age\_group in ('0-14', '15-24'); /\* For 2 selected age groups \*/

**run**;