```
/* Import the Excel file into SAS */
PROC IMPORT DATAFILE="C:\Users\nasir\OneDrive\Desktop\Multivariate data analysis (BIA 652)\Final_data.
            OUT=final_data
            DBMS=XLSX REPLACE;
RUN;
PROC PRINT DATA=final_data;
RUN;
/* Perform descriptive statistics on metric variables */
PROC MEANS DATA=final_data;
           VAR age_at_diagnosis incidence_year vit_stat_interval num_of_investigated_node num_of_posit
RUN;
/* Calculate the correlation matrix */
proc corr data=final_data outp=corr_matrix;
  var age_at_diagnosis incidence_year vit_stat_interval num_of_investigated_node num_of_positive_nodes
run;

/* Print the correlation matrix */
proc print data=corr_matrix;
run;
/* Perform correlation analysis */
proc corr data=final_data;
  var vit_stat_interval;
  with age_at_diagnosis incidence_year num_of_investigated_node num_of_positive_nodes tumor_size;
  by laterality _differentiation_grade er_status pr_status Pathological_tumor_stage Pathological_nodal_
run;
PROC FREQ DATA=final_data;
  TABLES vit_status laterality _differentiation_grade er_status pr_status
         Pathological_tumor_stage Pathological_nodal_stage her2_status
         multifocality Axillary_Node_Dissection immediate_reconstruction
         tumor_morphology surgery_type chemo_therapy hormonal_therapy
         radio_therapy targeted_therapy;
RUN;
/* Perform univariate analysis using PROC UNIVARIATE with box plots */
PROC UNIVARIATE DATA=final_data Normal Plot;
  VAR age_at_diagnosis incidence_year vit_stat_interval num_of_investigated_node
      num_of_positive_nodes tumor_size;
RUN;
* Logistic Regression;
Proc Logistic Data = final_data;
Model vit_status(event='0') = age_at_diagnosis incidence_year  vit_stat_interval laterality _different

/ Selection=Stepwise SLEntry=0.05 SLStay=0.05 Details
LackFit RSquare CTable PProb =(0 to 1 by .10);
run;
*;
*Final;
Proc Logistic Data = final_data OutModel=Logistic_final;
Model vit_status(event='0') = hormonal_therapy radio_therapy _differentiation_grade chemo_therapy num_
/ LackFit RSquare CTable PProb =(0.40 to 0.60 by .01);
run;
/* Create final_data60 dataset (60% of original data) */
Data final_data60;
    Set final_data;
    If _N_ <= 7359; /* 60% of 12266 is approximately 7359 */
run;
```

```
* Create Fianl_data40 dataset (40% of original data);
Data final_data40;
    Set final_data;
    If _N_ > 7359; * Selects remaining observations after the first 7359 (approximately 40%);
run;
Proc Print Data= final_data40;
run;
Proc Print Data= final_data60;
run;
* Logistic Regression of 60% data;
Proc Logistic Data = final_data60;
Model vit_status(event='0') = age_at_diagnosis incidence_year  vit_stat_interval laterality _different

/ Selection=Stepwise SLEntry=0.05 SLStay=0.05 Details
LackFit RSquare CTable PProb =(0 to 1 by .10);
run;
*Final 60%;
Proc Logistic Data = final_data60 OutModel=Logistic_final_60;
Model vit_status(event='0') = hormonal_therapy radio_therapy _differentiation_grade chemo_therapy num_
/ LackFit RSquare CTable PProb =(0.40 to 0.60 by .01);
run;
* Original Split60 Logistic Model Fitted to Split40 validation Data;
*;
Proc Logistic InModel=Logistic_final_60;
Score Data = final_data60 (Keep = vit_status hormonal_therapy radio_therapy _differentiation_grade che
run;
* Proc Freq Crosstabulations Original and Holdout Validation Datasets;
*;
Proc Print Data = final_data60Score;
Proc Freq Data = final_data60Score;
Table F_vit_status * I_vit_status;
*;
Proc Logistic InModel=Logistic_final_60;
    Score Data=final_data40 (Keep=vit_status hormonal_therapy radio_therapy
    _differentiation_grade chemo_therapy num_of_positive_nodes age_at_diagnosis
    er_status Pathological_tumor_stage multifocality Axillary_Node_Dissection)
    Out=final_data40Score;
Run;

Proc Print Data = final_data40Score;
Proc Freq Data = final_data40Score;
Table F_vit_status * I_vit_status;
run;
```