

```

libname Project '/home/u61892910/DS6371/Project';

/*****
***DS6371 Project
***Amy Adyanthaya & Nicholas Mueller
***2023-04-09
*****/

***Question 1***;
***Read in the data and filter on the three neighborhoods***;
proc sort data=Project.train out=A1(keep=Neighborhood id GrLivArea SalePrice);
    by Neighborhood GrLivArea SalePrice id;
    where Neighborhood in('Edwards', 'NAmes', 'BrkSide');
run;

***Check to see if there are any missing values in the three variables***;
***Create transformation variables***;
data A1;
    set A1;
    by Neighborhood GrLivArea SalePrice id;
    if missing(neighborhood) then missn = 1;
    if missing(grlivarea) then missg = 1;
    if missing(saleprice) then misss = 1;
    LGrLivArea=log(GrLivArea);
    LSalePrice=log(SalePrice);
run;

%macro NoMissing;
***Verify no missing value count for each variable
***No missing values were identified;
proc sql;
    select sum(missn) as Sum_MissN
    from work.A1;
    select sum(missg) as Sum_MissG
    from work.A1;
    select sum(misss) as Sum_MissS
    from work.A1;
quit;
%mend NoMissing;

***Plot the data to determine if any assumptions violated***;
ods graphics / attrpriority=none;
proc sgplot data=a1;
    styleattrs datasymbols = (trianglefilled squarefilled starfilled)
    datacontrastcolors = (red green blue);

```

```

scatter x=GrLivArea y=SalePrice / group=Neighborhood;
reg x=GrLivArea y=SalePrice / nomarkers;
xaxis label = "Grade Living Area (sqft)";
yaxis label = "Sale Price ($100)";
title1 "Home Sales Price by Grade Living Area";
title2 "Neighborhoods: Brookside, Edwards, Northwest Ames";

run;
quit;

***Run the GLM to check the model & include all plots***;
***Check the assumptions***;
proc glm data=A1 plots=all;
    class Neighborhood;
    *model SalePrice = GrLivArea|Neighborhood / solution;
    model SalePrice = GrLivArea Neighborhood / solution;
run;

***Plot the transformations on the data to identify the ideal model***;
%macro ReferenceOnly;
***Linear-Log***;
ods graphics / attrpriority=none;
proc sgplot data=a1;
    styleattrs datasymbols = (trianglefilled squarefilled starfilled)
        datacontrastcolors = (red green blue);
    scatter x=LGrLivArea y=SalePrice / group=Neighborhood;
    reg x=LGrLivArea y=SalePrice / nomarkers;
    xaxis label = "log(Grade Living Area (sqft))";
    yaxis label = "Sale Price ($100)";
    title1 "Linear-Log: Home Sales Price by Grade Living Area";
    title2 "Neighborhoods: Brookside, Edwards, Northwest Ames";

run;
quit;

***Log-Linear***;
ods graphics / attrpriority=none;
proc sgplot data=a1;
    styleattrs datasymbols = (trianglefilled squarefilled starfilled)
        datacontrastcolors = (red green blue);
    scatter x=GrLivArea y=LSalePrice / group=Neighborhood;
    reg x=GrLivArea y=LSalePrice / nomarkers;
    xaxis label = "Grade Living Area (sqft)";
    yaxis label = "log(Sale Price ($100))";
    title1 "Log-Linear: Home Sales Price by Grade Living Area";
    title2 "Neighborhoods: Brookside, Edwards, Northwest Ames";

run;
quit;

```

```

%mend ReferenceOnly;

***Log-Log***;
ods graphics / attrpriority=none;
proc sgplot data=a1;
    styleattrs datasymbols = (trianglefilled squarefilled starfilled)
        datacontrastcolors = (red green blue);
    scatter x=LGrLivArea y=LSalePrice / group=Neighborhood;
    reg x=LGrLivArea y=LSalePrice / nomarkers;
    xaxis label = "log(Grade Living Area (sqft))";
    yaxis label = "log(Sale Price ($100))";
    title1 "Log-Log: Home Sales Price by Grade Living Area";
    title2 "Neighborhoods: Brookside, Edwards, Northwest Ames";
run;
quit;

***Look at the log-log transformed data***;
ods graphics on;
proc glm data=A1(rename=(LGrLivArea='log(Grade Living Area (sqft))'n
                        LSalePrice='log(Sale Price ($100))'n)) plots=all;
    class Neighborhood;
    model 'log(Sale Price ($100))'n = 'log(Grade Living Area (sqft))'n Neighborhood / solution;
run;
quit;

***Obtain the estimates and analysis for the data***;
ods graphics on;
proc glmselect data=A1(rename=(LGrLivArea='log(Grade Living Area (sqft))'n
                        LSalePrice='log(Sale Price ($100))'n)) plots(stepaxis=number) = (criterionpanel ASEPlot);
    class Neighborhood / order=internal;
    model 'log(Sale Price ($100))'n = 'log(Grade Living Area (sqft))'n Neighborhood
        / selection=stepwise(choose=CV stop=CV) cvdetails=All showpvalues include=4;
run;
quit;

```