6371 Kaggle Project Appendix

Jodi Pafford, Tori Wheelis, Sara Zaheri

#Import data through file import and run proc print to verify appropriate file formats.

proc print data = Train;

run;

proc print data = Test;

run;

#Add a dot (.) in the dataset where there is no SalePrice (the response variable).

data Test;

set Test;

SalePrice = .;

run;

proc print data = Test;

run;

#Merge the two datasets (Test and Train) into “Train2” and proc print for confirmation.

data Train2;

set Train Test;

run;

proc print data = Train2;

run;

#Sort the file by neighborhood and edit the file to only include the neighborhoods of “Edwards”, “Names”, and “BrkSide”. Also removed other not needed fields. This file kept id, Neighborhood, GrLlvArea, and SalePrice. This file is called “Q1”

proc sort data = Train2 out=SrtTrain;

by Neighborhood;

run;

Data Q1;

set SrtTrain;

logGrLIvArea = log(GrLIvArea);

LogSalePrice = Log(SalePrice);

KEEP id Neighborhood GrLIvArea SalePrice;

if (Neighborhood = 'NAmes') or (Neighborhood = 'Edwards') or (Neighborhood = 'BrkSide');

run;

#Explore Different Transformations with the data

#Transformed SalePrice and grLivingArea using log function

Data Q1;

set Q1;

logGrLIvArea = log(GrLIvArea);

Figure 2
LogSalePrice = Log(SalePrice);

run;

proc sort data = Q1 out=Q1;

by id;

run;

#This becomes the final file with the two logged values added to the end of the dataset. Verify with proc print function (Figure 2).

proc print data = Q1;

run;

#For this question there are different combinations to be used for modeling the SalePrice:

#Model 1: no transformation, no outlier, and also no interaction.

proc glm data=Q1 plots=all;

class Neighborhood ;

model SalePrice = GrLIvArea Neighborhood / solution ;

output out = Results1 r=res p=pred student=stdres;

run;

#Model 2: Log transformation for SalePrice and no transformation for GrLivingAre, no outlier, and also no interaction (Log-Linear transformation model).

proc glm data=Q1 plots=all;

class Neighborhood ;

model LogSalePrice = GrLIvArea Neighborhood / solution ;

output out = Results3 r=res p=pred student=stdres;

run;

#Model 3: Log transformation for GrLivingAre and no transformation for SalePrice, no outlier, and also no interaction (Linear-Log transformation model).

proc glm data=Q1 plots=all;

class Neighborhood ;

model SalePrice = LogGrLIvArea Neighborhood / solution ;

output out = Results4 r=res p=pred student=stdres;

run;

#Model 4: Log transformation for both GrLivingAre and SalePrice, no outlier, and also no interaction (Linear-Log transformation model).

proc glm data=Q1 plots=all;

class Neighborhood ;

model LogSalePrice = LogGrLIvArea Neighborhood / solution ;

output out = Results2 r=res p=pred student=stdres;

run;

# The above results indicated possible outliers. Therefore, a search was done for residuals and looking for outliers.

Proc sort data = results2 out = rr;

by res;

proc print data=rr;

run;

#Eliminate outliers that were found

data Q1\_2;

set Q1;

if (id = 725) or (id = 411) then delete;

run;

\*no transformation and removed outlier = 2; (Figure 8)

proc glm data=Q1\_2 plots=all;

class Neighborhood ;

model SalePrice = GrLIvArea Neighborhood / solution ;

output out = Results1 r=res p=pred student=stdres;

run;

\*log-log model and removed outlier = 2; (Figure 9)

proc glm data=Q1\_2 plots=all;

class Neighborhood ;

model LogSalePrice = LogGrLIvArea Neighborhood / solution ;

output out = Results2 r=res p=pred student=stdres;

run;