## Analysis 1

FILENAME REFFILE '/folders/myfolders/sasuser.v94/test.csv';

PROC IMPORT DATAFILE=REFFILE

DBMS=CSV

OUT=test;

GETNAMES=YES;

RUN;

FILENAME REFFILE '/folders/myfolders/sasuser.v94/train.csv';

PROC IMPORT DATAFILE=REFFILE

DBMS=CSV

OUT=train;

GETNAMES=YES;

RUN;

/\*Filter only relevant Neighborhoods for Q1. NAmes, Edwards, BrkSide\*/

data neighborhoodSS;

set train;

where Neighborhood = 'NAmes' or Neighborhood = 'Edwards' or Neighborhood='BrkSide';

run;

/\* Create GrLivAreaper100 variable, log Transform the SalesPrice and GrLivAreaper100 to re-run model \*/

data logneighborhoodSS;

set neighborhoodSS;

GrLivAreaper100 = GrLivArea/100;

logSalePrice = log(SalePrice);

logGrLivAreaper100 = log(GrLivArea/100);

run;

/\*proc print data = logneighborhoodSS;

run;\*/

/\*Scatterplot the Saleprice to GrLivArea for the three requested neighborhoods\*/

proc sgplot data=logneighborhoodSS;

scatter x=GrLivAreaper100 y=SalePrice / group=Neighborhood;

run;

/\* Model for Salesprice to GrLivArea\*/

proc glm data = logneighborhoodSS plots=all;

class Neighborhood;

model SalePrice = GrLivAreaper100 | Neighborhood / cli clm clparm;

run;

/\*Filter out all homes greater than 40 100 sq.ft.\*/

data logneighborhoodSSsmall;

set logneighborhoodSS;

where GrLivAreaper100 <= 40;

run;

/\* Model for Salesprice to GrLivArea\*/

proc glm data = logneighborhoodSSsmall plots=all;

class Neighborhood;

model SalePrice = GrLivAreaper100 | Neighborhood / cli clm clparm;

run;

/\*Filter out all homes that costs more than $300,000\*/

data logneighborhoodSScheap;

set logneighborhoodSS;

where SalePrice <= 300000;

run;

/\* Model for Salesprice to GrLivArea\*/

proc glm data = logneighborhoodSScheap plots=all;

class Neighborhood;

model SalePrice = GrLivAreaper100 | Neighborhood / cli clm clparm;

run;

/\*Filter out all homes that costs more than $300,000 and smaller than 40 100 sq. ft.\*/

data logneighborhoodSScheapsmall;

set logneighborhoodSS;

where SalePrice <= 300000 and GrLivAreaper100 <= 40;

run;

/\* Model for Salesprice to GrLivArea\*/

proc glm data = logneighborhoodSScheapsmall plots=all;

class Neighborhood;

model SalePrice = GrLivAreaper100 | Neighborhood / cli clm clparm solution;

run;

/\* Model with log transformed SalePrice\*/

proc glm data = logneighborhoodSS plots=all;

class Neighborhood;

model logSalePrice = GrLivAreaper100 | Neighborhood / cli clm clparm;

run;

/\* Model with log transformed GrLivAreaper100\*/

proc glm data = logneighborhoodSS plots=all;

class Neighborhood;

model SalePrice = logGrLivAreaper100 | Neighborhood / cli clm clparm;

run;

/\* Model with log-log transformed SalePrice-GrLivAreaper100\*/

proc glm data = logneighborhoodSS plots=all;

class Neighborhood;

model logSalePrice = logGrLivAreaper100 | Neighborhood / cli clm clparm;

run;

/\* Model with log-log transformed SalePrice-GrLivAreaper100 on homes that are less than $300,000 and 40 100 sq. ft.\*/

proc glm data = logneighborhoodSScheapsmall plots=all;

class Neighborhood;

model logSalePrice = logGrLivAreaper100 | Neighborhood / cli clm clparm;

run;

## Analysis 2

#R code for EDA

#masterAN.csv is combination of train.csv and test.csv

trainDF <- read.csv("masterAN.csv", header = T, stringsAsFactors = F)

trainDF <- subset(trainDF[1:1470,])

trainDF$SalePrice <- as.integer(trainDF$SalePrice)

titles <- names(trainDF)

for (i in 1:length(trainDF)){

if (is.character(trainDF[[i]])){

boxplot(trainDF$SalePrice ~ trainDF[[i]], main = paste("Sale price vs ", titles[i], sep=""), ylab = titles[i], xlab = "Sale price", horizontal = T, las = 1)

}

else if (is.integer(trainDF[[i]])){

plot(x = trainDF[[i]], y = trainDF$SalePrice, main = paste("Sale price vs ", titles[i], sep=""), xlab = titles[i], ylab = "Sale price")

}

}

%\*SAS code for model selection (Forward, Backward, Stepwise, and custom user defined);

data train;

infile '/folders/myfolders/sasuser.v94/masterAN.csv' dlm=',' firstobs=2;

input Id MSSubClass MSZoning $ LotFrontage LotArea Street $ Alley $ LotShape $ LandContour $

Utilities $ LotConfig $ LandSlope $ Neighborhood $ Condition1 $ Condition2 $ BldgType $

HouseStyle $ OverallQual OverallCond YearBuilt YearRemodAdd RoofStyle $ RoofMatl $

Exterior1st $ Exterior2nd $ MasVnrType $ MasVnrArea ExterQual $ ExterCond $ Foundation $

BsmtQual $ BsmtCond $ BsmtExposure $ BsmtFinType1 $ BsmtFinSF1 BsmtFinType2 $ BsmtFinSF2 BsmtUnfSF TotalBsmtSF Heating $ HeatingQC $ CentralAir $ Electrical $ \_1stFlrSF \_2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr KitchenQual $ TotRmsAbvGrd Functional $ Fireplaces FireplaceQu $ GarageType $ GarageYrBlt GarageFinish $ GarageCars GarageArea GarageQual $ GarageCond $ PavedDrive $ WoodDeckSF OpenPorchSF EnclosedPorch \_3SsnPorch ScreenPorch PoolArea PoolQC $ Fence $ MiscFeature $ MiscVal MoSold YrSold SaleType $ SaleCondition $ SalePrice;

%\*Forward selection;

proc glmselect data=train seed=1;

class MSZoning Street Alley LotShape LandContour Utilities LotConfig LandSlope

Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl

Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual

BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir

Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish

GarageQual GarageCond PavedDrive Fence MiscFeature SaleType SaleCondition PoolQC;

model SalePrice=MSSubClass--SaleCondition / selection=forward (choose=CV

stop=CV) cvmethod=split(10) CVdetails;

output out=forward\_results p=Predict;

run;

data forward\_results2;

set forward\_results;

if Predict > 0 then

SalePrice=Predict;

if Predict < 0 then

SalePrice=10000;

keep id SalePrice;

where id > 1460;

run;

proc export data=forward\_results2

outfile='/folders/myfolders/sasuser.v94/submitForward1.csv' dbms=csv replace;

run;

proc print data=forward\_results2;

run;

%\*Backward elimination;

proc glmselect data=train seed=1;

class MSZoning Street Alley LotShape LandContour Utilities LotConfig LandSlope

Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl

Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual

BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir

Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish

GarageQual GarageCond PavedDrive Fence MiscFeature SaleType SaleCondition PoolQC;

model SalePrice=MSSubClass--SaleCondition / selection=backward(choose=CV

stop=CV) cvmethod=split(10) CVdetails;

output out=backward\_results p=Predict;

run;

data backward\_results2;

set backward\_results;

if Predict > 0 then

SalePrice=Predict;

if Predict < 0 then

SalePrice=10000;

keep id SalePrice;

where id > 1460;

run;

proc export data=backward\_results2

outfile='/folders/myfolders/sasuser.v94/submitBackward1.csv' dbms=csv replace;

run;

proc print data=backward\_results2;

run;

%\*Stepwise selection;

proc glmselect data=train seed=1;

class MSZoning Street Alley LotShape LandContour Utilities LotConfig LandSlope

Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl

Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual

BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir

Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish

GarageQual GarageCond PavedDrive Fence MiscFeature SaleType SaleCondition PoolQC;

model SalePrice=MSSubClass--SaleCondition / selection=stepwise(choose=CV

stop=CV) cvmethod=split(10) CVdetails;

output out=stepwise\_results p=Predict;

run;

data stepwise\_results2;

set stepwise\_results;

if Predict > 0 then

SalePrice=Predict;

if Predict < 0 then

SalePrice=10000;

keep id SalePrice;

where id > 1460;

run;

proc export data=stepwise\_results2

outfile='/folders/myfolders/sasuser.v94/submitStepwise1.csv' dbms=csv replace;

run;

proc print data=stepwise\_results2;

run;

%\*Custom model;

proc glmselect data=train2 plots=all;

class KitchenQual GarageFinish BsmtQual ExterQual MasVnrType Neighborhood;

model SalePrice=LotFrontage LotArea OverallQual OverallCond YearBuilt

BsmtFinSF1 TotalBsmtSF \_1stFlrSF \_2ndFlrSF GrLivArea FullBath TotRmsAbvGrd

GarageYrBlt GarageCars GarageArea WoodDeckSF OpenPorchSF KitchenQual

GarageFinish BsmtQual ExterQual MasVnrType Neighborhood / selection=none;

output out=results p=Predict;

run;

data results3;

set results;

if Predict > 0 then

SalePrice=Predict;

if Predict < 0 then

SalePrice=10000;

keep id SalePrice;

where id > 1460;

run;

proc export data=results3 outfile='/folders/myfolders/sasuser.v94/submitCustom2.csv'

dbms=csv replace;

run;