

MTH404 R Project

Mustafif Khan

Loading Required Packages

Before the dataset can be analyzed, the following packages must be imported:

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
## v dplyr      1.1.2      v readr      2.1.4
## v forcats    1.0.0      v stringr   1.5.0
## v ggplot2    3.4.2      v tibble    3.2.1
## v lubridate  1.9.2      v tidyr     1.3.0
## v purrr      1.0.1
```

```
## -- Conflicts ----- tidyverse_conflicts() --
```

```
## x dplyr::filter() masks stats::filter()
```

```
## x dplyr::lag()     masks stats::lag()
```

```
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(ggplot2)
```

```
library(lubridate)
```

```
library(gridExtra)
```

```
##
```

```
## Attaching package: 'gridExtra'
```

```
##
```

```
## The following object is masked from 'package:dplyr':
```

```
##
```

```
##      combine
```

```
library(caTools)
```

```
library(dplyr)
```

Reading the Training and Test Data

From the kaggle competition, we are provided the CSV data for training and testing, and the dataset is imported like the following:

```
train_data <- read.csv("train.csv")
```

```
test_data <- read.csv("test.csv")
```

```
head(train_data)
```

```
##   Id MSSubClass MSZoning LotFrontage LotArea Street Alley LotShape LandContour
## 1  1           60      RL           65    8450   Pave  <NA>      Reg          Lvl
```

##	2	2	20	RL	80	9600	Pave	<NA>	Reg	Lvl
##	3	3	60	RL	68	11250	Pave	<NA>	IR1	Lvl
##	4	4	70	RL	60	9550	Pave	<NA>	IR1	Lvl
##	5	5	60	RL	84	14260	Pave	<NA>	IR1	Lvl
##	6	6	50	RL	85	14115	Pave	<NA>	IR1	Lvl
##	Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType									
##	1	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam		
##	2	AllPub	FR2	Gtl	Veenker	Feedr	Norm	1Fam		
##	3	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam		
##	4	AllPub	Corner	Gtl	Crawfor	Norm	Norm	1Fam		
##	5	AllPub	FR2	Gtl	NoRidge	Norm	Norm	1Fam		
##	6	AllPub	Inside	Gtl	Mitchel	Norm	Norm	1Fam		
##	HouseStyle OverallQual OverallCond YearBuilt YearRemodAdd RoofStyle RoofMatl									
##	1	2Story		7	5	2003	2003	Gable	CompShg	
##	2	1Story		6	8	1976	1976	Gable	CompShg	
##	3	2Story		7	5	2001	2002	Gable	CompShg	
##	4	2Story		7	5	1915	1970	Gable	CompShg	
##	5	2Story		8	5	2000	2000	Gable	CompShg	
##	6	1.5Fin		5	5	1993	1995	Gable	CompShg	
##	Exterior1st Exterior2nd MasVnrType MasVnrArea ExterQual ExterCond Foundation									
##	1	VinylSd	VinylSd	BrkFace	196	Gd	TA	PConc		
##	2	MetalSd	MetalSd	None	0	TA	TA	CBlock		
##	3	VinylSd	VinylSd	BrkFace	162	Gd	TA	PConc		
##	4	Wd Sdng	Wd Shng	None	0	TA	TA	BrkTil		
##	5	VinylSd	VinylSd	BrkFace	350	Gd	TA	PConc		
##	6	VinylSd	VinylSd	None	0	TA	TA	Wood		
##	BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinSF1 BsmtFinType2									
##	1	Gd	TA	No	GLQ	706	Unf			
##	2	Gd	TA	Gd	ALQ	978	Unf			
##	3	Gd	TA	Mn	GLQ	486	Unf			
##	4	TA	Gd	No	ALQ	216	Unf			
##	5	Gd	TA	Av	GLQ	655	Unf			
##	6	Gd	TA	No	GLQ	732	Unf			
##	BsmtFinSF2 BsmtUnfSF TotalBsmtSF Heating HeatingQC CentralAir Electrical									
##	1	0	150	856	GasA	Ex	Y	SBrkr		
##	2	0	284	1262	GasA	Ex	Y	SBrkr		
##	3	0	434	920	GasA	Ex	Y	SBrkr		
##	4	0	540	756	GasA	Gd	Y	SBrkr		
##	5	0	490	1145	GasA	Ex	Y	SBrkr		
##	6	0	64	796	GasA	Ex	Y	SBrkr		
##	X1stFlrSF X2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath									
##	1	856	854	0	1710	1	0	2		
##	2	1262	0	0	1262	0	1	2		
##	3	920	866	0	1786	1	0	2		
##	4	961	756	0	1717	1	0	1		
##	5	1145	1053	0	2198	1	0	2		
##	6	796	566	0	1362	1	0	1		
##	HalfBath BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd Functional									
##	1	1	3	1	Gd	8	Typ			
##	2	0	3	1	TA	6	Typ			
##	3	1	3	1	Gd	6	Typ			
##	4	0	3	1	Gd	7	Typ			
##	5	1	4	1	Gd	9	Typ			
##	6	1	1	1	TA	5	Typ			

```
##      Fireplaces FireplaceQu GarageType GarageYrBlt GarageFinish GarageCars
## 1           0          <NA>    Attchd      2003          RFn          2
## 2           1           TA    Attchd      1976          RFn          2
## 3           1           TA    Attchd      2001          RFn          2
## 4           1           Gd    Detchd      1998          Unf          3
## 5           1           TA    Attchd      2000          RFn          3
## 6           0          <NA>    Attchd      1993          Unf          2
##      GarageArea GarageQual GarageCond PavedDrive WoodDeckSF OpenPorchSF
## 1          548          TA          TA          Y          0          61
## 2          460          TA          TA          Y        298          0
## 3          608          TA          TA          Y          0          42
## 4          642          TA          TA          Y          0          35
## 5          836          TA          TA          Y        192          84
## 6          480          TA          TA          Y         40          30
##      EnclosedPorch X3SsnPorch ScreenPorch PoolArea PoolQC Fence MiscFeature
## 1              0            0            0          0 <NA> <NA>    <NA>
## 2              0            0            0          0 <NA> <NA>    <NA>
## 3              0            0            0          0 <NA> <NA>    <NA>
## 4             272            0            0          0 <NA> <NA>    <NA>
## 5              0            0            0          0 <NA> <NA>    <NA>
## 6              0           320            0          0 <NA> MnPrv    Shed
##      MiscVal MoSold YrSold SaleType SaleCondition SalePrice
## 1          0         2   2008         WD        Normal    208500
## 2          0         5   2007         WD        Normal    181500
## 3          0         9   2008         WD        Normal    223500
## 4          0         2   2006         WD      Abnorml    140000
## 5          0        12   2008         WD        Normal    250000
## 6         700        10   2009         WD        Normal    143000
```

```
head(test_data)
```

```
##      Id MSSubClass MSZoning LotFrontage LotArea Street Alley LotShape
## 1 1461          20      RH          80   11622   Pave <NA>    Reg
## 2 1462          20      RL          81   14267   Pave <NA>    IR1
## 3 1463          60      RL          74   13830   Pave <NA>    IR1
## 4 1464          60      RL          78   9978    Pave <NA>    IR1
## 5 1465         120      RL          43   5005    Pave <NA>    IR1
## 6 1466          60      RL          75  10000    Pave <NA>    IR1
##      LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2
## 1          Lvl1    AllPub    Inside      Gtl      Names      Feedr      Norm
## 2          Lvl1    AllPub    Corner      Gtl      Names      Norm      Norm
## 3          Lvl1    AllPub    Inside      Gtl    Gilbert      Norm      Norm
## 4          Lvl1    AllPub    Inside      Gtl    Gilbert      Norm      Norm
## 5          HLS    AllPub    Inside      Gtl    StoneBr      Norm      Norm
## 6          Lvl1    AllPub    Corner      Gtl    Gilbert      Norm      Norm
##      BldgType HouseStyle OverallQual OverallCond YearBuilt YearRemodAdd RoofStyle
## 1      1Fam      1Story          5          6     1961      1961      Gable
## 2      1Fam      1Story          6          6     1958      1958      Hip
## 3      1Fam      2Story          5          5     1997      1998      Gable
## 4      1Fam      2Story          6          6     1998      1998      Gable
## 5    TwnhsE      1Story          8          5     1992      1992      Gable
## 6      1Fam      2Story          6          5     1993      1994      Gable
##      RoofMatl Exterior1st Exterior2nd MasVnrType MasVnrArea ExterQual ExterCond
## 1    CompShg    VinylSd    VinylSd      None          0          TA          TA
## 2    CompShg      Wd Sdng      Wd Sdng    BrkFace        108          TA          TA
```

## 3	CompShg	VinylSd	VinylSd	None	0	TA	TA
## 4	CompShg	VinylSd	VinylSd	BrkFace	20	TA	TA
## 5	CompShg	HdBoard	HdBoard	None	0	Gd	TA
## 6	CompShg	HdBoard	HdBoard	None	0	TA	TA
##	Foundation	BsmtQual	BsmtCond	BsmtExposure	BsmtFinType1	BsmtFinSF1	
## 1	CBlock	TA	TA	No	Rec	468	
## 2	CBlock	TA	TA	No	ALQ	923	
## 3	PConc	Gd	TA	No	GLQ	791	
## 4	PConc	TA	TA	No	GLQ	602	
## 5	PConc	Gd	TA	No	ALQ	263	
## 6	PConc	Gd	TA	No	Unf	0	
##	BsmtFinType2	BsmtFinSF2	BsmtUnfSF	TotalBsmtSF	Heating	HeatingQC	CentralAir
## 1	LwQ	144	270	882	GasA	TA	Y
## 2	Unf	0	406	1329	GasA	TA	Y
## 3	Unf	0	137	928	GasA	Gd	Y
## 4	Unf	0	324	926	GasA	Ex	Y
## 5	Unf	0	1017	1280	GasA	Ex	Y
## 6	Unf	0	763	763	GasA	Gd	Y
##	Electrical	X1stFlrSF	X2ndFlrSF	LowQualFinSF	GrLivArea	BsmtFullBath	
## 1	SBrkr	896	0	0	896	0	
## 2	SBrkr	1329	0	0	1329	0	
## 3	SBrkr	928	701	0	1629	0	
## 4	SBrkr	926	678	0	1604	0	
## 5	SBrkr	1280	0	0	1280	0	
## 6	SBrkr	763	892	0	1655	0	
##	BsmtHalfBath	FullBath	HalfBath	BedroomAbvGr	KitchenAbvGr	KitchenQual	
## 1	0	1	0	2	1	TA	
## 2	0	1	1	3	1	Gd	
## 3	0	2	1	3	1	TA	
## 4	0	2	1	3	1	Gd	
## 5	0	2	0	2	1	Gd	
## 6	0	2	1	3	1	TA	
##	TotRmsAbvGrd	Functional	Fireplaces	FireplaceQu	GarageType	GarageYrBlt	
## 1	5	Typ	0	<NA>	Attchd	1961	
## 2	6	Typ	0	<NA>	Attchd	1958	
## 3	6	Typ	1	TA	Attchd	1997	
## 4	7	Typ	1	Gd	Attchd	1998	
## 5	5	Typ	0	<NA>	Attchd	1992	
## 6	7	Typ	1	TA	Attchd	1993	
##	GarageFinish	GarageCars	GarageArea	GarageQual	GarageCond	PavedDrive	
## 1	Unf	1	730	TA	TA	Y	
## 2	Unf	1	312	TA	TA	Y	
## 3	Fin	2	482	TA	TA	Y	
## 4	Fin	2	470	TA	TA	Y	
## 5	RFn	2	506	TA	TA	Y	
## 6	Fin	2	440	TA	TA	Y	
##	WoodDeckSF	OpenPorchSF	EnclosedPorch	X3SsnPorch	ScreenPorch	PoolArea	PoolQC
## 1	140	0	0	0	120	0	<NA>
## 2	393	36	0	0	0	0	<NA>
## 3	212	34	0	0	0	0	<NA>
## 4	360	36	0	0	0	0	<NA>
## 5	0	82	0	0	144	0	<NA>
## 6	157	84	0	0	0	0	<NA>
##	Fence	MiscFeature	MiscVal	MoSold	YrSold	SaleType	SaleCondition

```
## 1 MnPrv      <NA>      0      6  2010      WD      Normal
## 2  <NA>      Gar2    12500      6  2010      WD      Normal
## 3 MnPrv      <NA>      0      3  2010      WD      Normal
## 4  <NA>      <NA>      0      6  2010      WD      Normal
## 5  <NA>      <NA>      0      1  2010      WD      Normal
## 6  <NA>      <NA>      0      4  2010      WD      Normal
```

We will now convert all the character columns to factors as shown below:

```
train_data <- as.data.frame(unclass(train_data), stringsAsFactors = TRUE)
test_data <- as.data.frame(unclass(test_data), stringsAsFactors = TRUE)
```

With the data converted, we can check if we are missing values in our training data:

```
# get the missing training values
missing_training_values <- sapply(train_data, function(x) sum(is.na(x)))
# create a data frame with column names and missing values
mtv_df <- data.frame(column = names(missing_training_values),
                     missing_values = missing_training_values)
mtv_df
```

```
##              column missing_values
## Id              Id              0
## MSSubClass      MSSubClass        0
## MSZoning        MSZoning          0
## LotFrontage     LotFrontage      259
## LotArea         LotArea          0
## Street          Street            0
## Alley          Alley            1369
## LotShape        LotShape          0
## LandContour     LandContour        0
## Utilities       Utilities          0
## LotConfig       LotConfig          0
## LandSlope       LandSlope          0
## Neighborhood    Neighborhood        0
## Condition1      Condition1          0
## Condition2      Condition2          0
## BldgType        BldgType          0
## HouseStyle      HouseStyle          0
## OverallQual     OverallQual          0
## OverallCond     OverallCond          0
## YearBuilt       YearBuilt          0
## YearRemodAdd    YearRemodAdd          0
## RoofStyle       RoofStyle          0
## RoofMatl        RoofMatl          0
## Exterior1st     Exterior1st          0
## Exterior2nd     Exterior2nd          0
## MasVnrType      MasVnrType          8
## MasVnrArea      MasVnrArea          8
## ExterQual       ExterQual          0
## ExterCond       ExterCond          0
## Foundation      Foundation          0
## BsmtQual        BsmtQual          37
## BsmtCond        BsmtCond          37
## BsmtExposure    BsmtExposure        38
## BsmtFinType1    BsmtFinType1        37
```

## BsmtFinSF1	BsmtFinSF1	0
## BsmtFinType2	BsmtFinType2	38
## BsmtFinSF2	BsmtFinSF2	0
## BsmtUnfSF	BsmtUnfSF	0
## TotalBsmtSF	TotalBsmtSF	0
## Heating	Heating	0
## HeatingQC	HeatingQC	0
## CentralAir	CentralAir	0
## Electrical	Electrical	1
## X1stFlrSF	X1stFlrSF	0
## X2ndFlrSF	X2ndFlrSF	0
## LowQualFinSF	LowQualFinSF	0
## GrLivArea	GrLivArea	0
## BsmtFullBath	BsmtFullBath	0
## BsmtHalfBath	BsmtHalfBath	0
## FullBath	FullBath	0
## HalfBath	HalfBath	0
## BedroomAbvGr	BedroomAbvGr	0
## KitchenAbvGr	KitchenAbvGr	0
## KitchenQual	KitchenQual	0
## TotRmsAbvGrd	TotRmsAbvGrd	0
## Functional	Functional	0
## Fireplaces	Fireplaces	0
## FireplaceQu	FireplaceQu	690
## GarageType	GarageType	81
## GarageYrBlt	GarageYrBlt	81
## GarageFinish	GarageFinish	81
## GarageCars	GarageCars	0
## GarageArea	GarageArea	0
## GarageQual	GarageQual	81
## GarageCond	GarageCond	81
## PavedDrive	PavedDrive	0
## WoodDeckSF	WoodDeckSF	0
## OpenPorchSF	OpenPorchSF	0
## EnclosedPorch	EnclosedPorch	0
## X3SsnPorch	X3SsnPorch	0
## ScreenPorch	ScreenPorch	0
## PoolArea	PoolArea	0
## PoolQC	PoolQC	1453
## Fence	Fence	1179
## MiscFeature	MiscFeature	1406
## MiscVal	MiscVal	0
## MoSold	MoSold	0
## YrSold	YrSold	0
## SaleType	SaleType	0
## SaleCondition	SaleCondition	0
## SalePrice	SalePrice	0

As you can notice there are data in this dataset that contains large amounts of missing values, such as columns like PoolQC, Fence, etc. We will be removing the columns with missing values like the following:

```
data.train <- subset(train_data, select = -c(Id, LotFrontage, Alley, MasVnrType, MasVnrArea, BsmtQual,
head(data.train)
```

```
## MSSubClass MSZoning LotArea Street LotShape LandContour Utilities LotConfig
```

## 1	60	RL	8450	Pave	Reg	Lvl	AllPub	Inside		
## 2	20	RL	9600	Pave	Reg	Lvl	AllPub	FR2		
## 3	60	RL	11250	Pave	IR1	Lvl	AllPub	Inside		
## 4	70	RL	9550	Pave	IR1	Lvl	AllPub	Corner		
## 5	60	RL	14260	Pave	IR1	Lvl	AllPub	FR2		
## 6	50	RL	14115	Pave	IR1	Lvl	AllPub	Inside		
##	LandSlope	Neighborhood	Condition1	Condition2	BldgType	HouseStyle	OverallQual			
## 1	Gtl	CollgCr	Norm	Norm	1Fam	2Story	7			
## 2	Gtl	Veenker	Feedr	Norm	1Fam	1Story	6			
## 3	Gtl	CollgCr	Norm	Norm	1Fam	2Story	7			
## 4	Gtl	Crawfor	Norm	Norm	1Fam	2Story	7			
## 5	Gtl	NoRidge	Norm	Norm	1Fam	2Story	8			
## 6	Gtl	Mitchel	Norm	Norm	1Fam	1.5Fin	5			
##	OverallCond	YearBuilt	YearRemodAdd	RoofStyle	RoofMatl	Exterior1st	Exterior2nd			
## 1	5	2003	2003	Gable	CompShg	VinylSd	VinylSd			
## 2	8	1976	1976	Gable	CompShg	MetalSd	MetalSd			
## 3	5	2001	2002	Gable	CompShg	VinylSd	VinylSd			
## 4	5	1915	1970	Gable	CompShg	Wd Sdng	Wd Shng			
## 5	5	2000	2000	Gable	CompShg	VinylSd	VinylSd			
## 6	5	1993	1995	Gable	CompShg	VinylSd	VinylSd			
##	ExterQual	ExterCond	Foundation	BsmtFinSF1	BsmtFinSF2	BsmtUnfSF	TotalBsmtSF			
## 1	Gd	TA	PConc	706	0	150	856			
## 2	TA	TA	CBlock	978	0	284	1262			
## 3	Gd	TA	PConc	486	0	434	920			
## 4	TA	TA	BrkTil	216	0	540	756			
## 5	Gd	TA	PConc	655	0	490	1145			
## 6	TA	TA	Wood	732	0	64	796			
##	Heating	HeatingQC	CentralAir	X1stFlrSF	X2ndFlrSF	LowQualFinSF	GrLivArea			
## 1	GasA	Ex	Y	856	854	0	1710			
## 2	GasA	Ex	Y	1262	0	0	1262			
## 3	GasA	Ex	Y	920	866	0	1786			
## 4	GasA	Gd	Y	961	756	0	1717			
## 5	GasA	Ex	Y	1145	1053	0	2198			
## 6	GasA	Ex	Y	796	566	0	1362			
##	BsmtFullBath	BsmtHalfBath	FullBath	HalfBath	BedroomAbvGr	KitchenAbvGr				
## 1	1	0	2	1	3	1				
## 2	0	1	2	0	3	1				
## 3	1	0	2	1	3	1				
## 4	1	0	1	0	3	1				
## 5	1	0	2	1	4	1				
## 6	1	0	1	1	1	1				
##	KitchenQual	TotRmsAbvGrd	Functional	Fireplaces	GarageCars	GarageArea				
## 1	Gd	8	Typ	0	2	548				
## 2	TA	6	Typ	1	2	460				
## 3	Gd	6	Typ	1	2	608				
## 4	Gd	7	Typ	1	3	642				
## 5	Gd	9	Typ	1	3	836				
## 6	TA	5	Typ	0	2	480				
##	PavedDrive	WoodDeckSF	OpenPorchSF	EnclosedPorch	X3SsnPorch	ScreenPorch				
## 1	Y	0	61	0	0	0				
## 2	Y	298	0	0	0	0				
## 3	Y	0	42	0	0	0				
## 4	Y	0	35	272	0	0				
## 5	Y	192	84	0	0	0				

```
## 6      Y      40      30      0      320      0
##   PoolArea MiscVal MoSold YrSold SaleType SaleCondition SalePrice
## 1      0      0      2   2008      WD      Normal      208500
## 2      0      0      5   2007      WD      Normal      181500
## 3      0      0      9   2008      WD      Normal      223500
## 4      0      0      2   2006      WD      Abnorml      140000
## 5      0      0     12   2008      WD      Normal      250000
## 6      0     700     10   2009      WD      Normal      143000
```

Now it would be good to also remove these columns from our testing data as well:

```
data.test <- subset(test_data, select = -c(Id, LotFrontage, Alley, MasVnrType, MasVnrArea, BsmtQual, Bsr
head(data.test)
```

```
##   MSSubClass MSZoning LotArea Street LotShape LandContour Utilities LotConfig
## 1      20      RH   11622   Pave      Reg      Lvl      AllPub      Inside
## 2      20      RL   14267   Pave      IR1      Lvl      AllPub      Corner
## 3      60      RL   13830   Pave      IR1      Lvl      AllPub      Inside
## 4      60      RL    9978   Pave      IR1      Lvl      AllPub      Inside
## 5     120      RL    5005   Pave      IR1      HLS      AllPub      Inside
## 6      60      RL   10000   Pave      IR1      Lvl      AllPub      Corner
##   LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle OverallQual
## 1      Gtl      Names      Feedr      Norm      1Fam      1Story      5
## 2      Gtl      Names      Norm      Norm      1Fam      1Story      6
## 3      Gtl      Gilbert      Norm      Norm      1Fam      2Story      5
## 4      Gtl      Gilbert      Norm      Norm      1Fam      2Story      6
## 5      Gtl      StoneBr      Norm      Norm      TwnhsE      1Story      8
## 6      Gtl      Gilbert      Norm      Norm      1Fam      2Story      6
##   OverallCond YearBuilt YearRemodAdd RoofStyle RoofMatl Exterior1st Exterior2nd
## 1      6      1961      1961      Gable      CompShg      VinylSd      VinylSd
## 2      6      1958      1958      Hip      CompShg      Wd Sdng      Wd Sdng
## 3      5      1997      1998      Gable      CompShg      VinylSd      VinylSd
## 4      6      1998      1998      Gable      CompShg      VinylSd      VinylSd
## 5      5      1992      1992      Gable      CompShg      HdBoard      HdBoard
## 6      5      1993      1994      Gable      CompShg      HdBoard      HdBoard
##   ExterQual ExterCond Foundation BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF
## 1      TA      TA      CBlocc      468      144      270      882
## 2      TA      TA      CBlocc      923      0      406      1329
## 3      TA      TA      PConc      791      0      137      928
## 4      TA      TA      PConc      602      0      324      926
## 5      Gd      TA      PConc      263      0      1017     1280
## 6      TA      TA      PConc      0      0      763      763
##   Heating HeatingQC CentralAir X1stFlrSF X2ndFlrSF LowQualFinSF GrLivArea
## 1      GasA      TA      Y      896      0      0      896
## 2      GasA      TA      Y     1329      0      0     1329
## 3      GasA      Gd      Y      928     701      0     1629
## 4      GasA      Ex      Y      926     678      0     1604
## 5      GasA      Ex      Y     1280      0      0     1280
## 6      GasA      Gd      Y      763     892      0     1655
##   BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr
## 1      0      0      1      0      2      1
## 2      0      0      1      1      3      1
## 3      0      0      2      1      3      1
## 4      0      0      2      1      3      1
## 5      0      0      2      0      2      1
```



```
## 6      0      0      2      1      3      1
## KitchenQual TotRmsAbvGrd Functional Fireplaces GarageCars GarageArea
## 1      TA      5      Typ      0      1      730
## 2      Gd      6      Typ      0      1      312
## 3      TA      6      Typ      1      2      482
## 4      Gd      7      Typ      1      2      470
## 5      Gd      5      Typ      0      2      506
## 6      TA      7      Typ      1      2      440
## PavedDrive WoodDeckSF OpenPorchSF EnclosedPorch X3SsnPorch ScreenPorch
## 1      Y      140      0      0      0      120
## 2      Y      393      36      0      0      0
## 3      Y      212      34      0      0      0
## 4      Y      360      36      0      0      0
## 5      Y      0      82      0      0      144
## 6      Y      157      84      0      0      0
## PoolArea MiscVal MoSold YrSold SaleType SaleCondition
## 1      0      0      6  2010      WD      Normal
## 2      0  12500      6  2010      WD      Normal
## 3      0      0      3  2010      WD      Normal
## 4      0      0      6  2010      WD      Normal
## 5      0      0      1  2010      WD      Normal
## 6      0      0      4  2010      WD      Normal
```

Analysis of the Train Data

Before we create a model to analyze our training data, we will need to look at the structure of our different columns using the `str()` function and get a summary of our training data.

```
str(train_data)

## 'data.frame':    1460 obs. of  81 variables:
## $ Id      : int  1 2 3 4 5 6 7 8 9 10 ...
## $ MSSubClass : int  60 20 60 70 60 50 20 60 50 190 ...
## $ MSZoning   : Factor w/ 5 levels "C (all)","FV",...: 4 4 4 4 4 4 4 4 5 4 ...
## $ LotFrontage : int  65 80 68 60 84 85 75 NA 51 50 ...
## $ LotArea     : int  8450 9600 11250 9550 14260 14115 10084 10382 6120 7420 ...
## $ Street      : Factor w/ 2 levels "Grvl","Pave": 2 2 2 2 2 2 2 2 2 2 ...
## $ Alley       : Factor w/ 2 levels "Grvl","Pave": NA NA NA NA NA NA NA NA NA ...
## $ LotShape    : Factor w/ 4 levels "IR1","IR2","IR3",...: 4 4 1 1 1 1 4 1 4 4 ...
## $ LandContour : Factor w/ 4 levels "Bnk","HLS","Low",...: 4 4 4 4 4 4 4 4 4 4 ...
## $ Utilities   : Factor w/ 2 levels "AllPub","NoSeWa": 1 1 1 1 1 1 1 1 1 1 ...
## $ LotConfig   : Factor w/ 5 levels "Corner","CulDSac",...: 5 3 5 1 3 5 5 1 5 1 ...
## $ LandSlope   : Factor w/ 3 levels "Gtl","Mod","Sev": 1 1 1 1 1 1 1 1 1 1 ...
## $ Neighborhood : Factor w/ 25 levels "Blmngtn","Blueste",...: 6 25 6 7 14 12 21 17 18 4 ...
## $ Condition1  : Factor w/ 9 levels "Artery","Feedr",...: 3 2 3 3 3 3 3 5 1 1 ...
## $ Condition2  : Factor w/ 8 levels "Artery","Feedr",...: 3 3 3 3 3 3 3 3 1 ...
## $ BldgType    : Factor w/ 5 levels "1Fam","2fmCon",...: 1 1 1 1 1 1 1 1 1 2 ...
## $ HouseStyle  : Factor w/ 8 levels "1.5Fin","1.5Unf",...: 6 3 6 6 6 1 3 6 1 2 ...
## $ OverallQual : int  7 6 7 7 8 5 8 7 7 5 ...
## $ OverallCond : int  5 8 5 5 5 5 5 6 5 6 ...
## $ YearBuilt   : int  2003 1976 2001 1915 2000 1993 2004 1973 1931 1939 ...
## $ YearRemodAdd : int  2003 1976 2002 1970 2000 1995 2005 1973 1950 1950 ...
## $ RoofStyle   : Factor w/ 6 levels "Flat","Gable",...: 2 2 2 2 2 2 2 2 2 2 ...
## $ RoofMatl    : Factor w/ 8 levels "ClyTile","CompShg",...: 2 2 2 2 2 2 2 2 2 2 ...
```

```

## $ Exterior1st : Factor w/ 15 levels "AsbShng","AsphShn",...: 13 9 13 14 13 13 13 7 4 9 ...
## $ Exterior2nd : Factor w/ 16 levels "AsbShng","AsphShn",...: 14 9 14 16 14 14 14 7 16 9 ...
## $ MasVnrType : Factor w/ 4 levels "BrkCmn","BrkFace",...: 2 3 2 3 2 3 4 4 3 3 ...
## $ MasVnrArea : int 196 0 162 0 350 0 186 240 0 0 ...
## $ ExterQual : Factor w/ 4 levels "Ex","Fa","Gd",...: 3 4 3 4 3 4 3 4 4 4 ...
## $ ExterCond : Factor w/ 5 levels "Ex","Fa","Gd",...: 5 5 5 5 5 5 5 5 5 5 ...
## $ Foundation : Factor w/ 6 levels "BrkTil","CBlock",...: 3 2 3 1 3 6 3 2 1 1 ...
## $ BsmtQual : Factor w/ 4 levels "Ex","Fa","Gd",...: 3 3 3 4 3 3 1 3 4 4 ...
## $ BsmtCond : Factor w/ 4 levels "Fa","Gd","Po",...: 4 4 4 2 4 4 4 4 4 4 ...
## $ BsmtExposure : Factor w/ 4 levels "Av","Gd","Mn",...: 4 2 3 4 1 4 1 3 4 4 ...
## $ BsmtFinType1 : Factor w/ 6 levels "ALQ","BLQ","GLQ",...: 3 1 3 1 3 3 3 1 6 3 ...
## $ BsmtFinSF1 : int 706 978 486 216 655 732 1369 859 0 851 ...
## $ BsmtFinType2 : Factor w/ 6 levels "ALQ","BLQ","GLQ",...: 6 6 6 6 6 6 6 2 6 6 ...
## $ BsmtFinSF2 : int 0 0 0 0 0 0 0 32 0 0 ...
## $ BsmtUnfSF : int 150 284 434 540 490 64 317 216 952 140 ...
## $ TotalBsmtSF : int 856 1262 920 756 1145 796 1686 1107 952 991 ...
## $ Heating : Factor w/ 6 levels "Floor","GasA",...: 2 2 2 2 2 2 2 2 2 2 ...
## $ HeatingQC : Factor w/ 5 levels "Ex","Fa","Gd",...: 1 1 1 3 1 1 1 1 1 3 ...
## $ CentralAir : Factor w/ 2 levels "N","Y": 2 2 2 2 2 2 2 2 2 2 ...
## $ Electrical : Factor w/ 5 levels "FuseA","FuseF",...: 5 5 5 5 5 5 5 5 5 2 ...
## $ X1stFlrSF : int 856 1262 920 961 1145 796 1694 1107 1022 1077 ...
## $ X2ndFlrSF : int 854 0 866 756 1053 566 0 983 752 0 ...
## $ LowQualFinSF : int 0 0 0 0 0 0 0 0 0 0 ...
## $ GrLivArea : int 1710 1262 1786 1717 2198 1362 1694 2090 1774 1077 ...
## $ BsmtFullBath : int 1 0 1 1 1 1 1 1 0 1 ...
## $ BsmtHalfBath : int 0 1 0 0 0 0 0 0 0 0 ...
## $ FullBath : int 2 2 2 1 2 1 2 2 2 1 ...
## $ HalfBath : int 1 0 1 0 1 1 0 1 0 0 ...
## $ BedroomAbvGr : int 3 3 3 3 4 1 3 3 2 2 ...
## $ KitchenAbvGr : int 1 1 1 1 1 1 1 1 2 2 ...
## $ KitchenQual : Factor w/ 4 levels "Ex","Fa","Gd",...: 3 4 3 3 3 4 3 4 4 4 ...
## $ TotRmsAbvGrd : int 8 6 6 7 9 5 7 7 8 5 ...
## $ Functional : Factor w/ 7 levels "Maj1","Maj2",...: 7 7 7 7 7 7 7 3 7 ...
## $ Fireplaces : int 0 1 1 1 1 0 1 2 2 2 ...
## $ FireplaceQu : Factor w/ 5 levels "Ex","Fa","Gd",...: NA 5 5 3 5 NA 3 5 5 5 ...
## $ GarageType : Factor w/ 6 levels "2Types","Attchd",...: 2 2 2 6 2 2 2 2 6 2 ...
## $ GarageYrBlt : int 2003 1976 2001 1998 2000 1993 2004 1973 1931 1939 ...
## $ GarageFinish : Factor w/ 3 levels "Fin","RFn","Unf": 2 2 2 3 2 3 2 2 3 2 ...
## $ GarageCars : int 2 2 2 3 3 2 2 2 2 1 ...
## $ GarageArea : int 548 460 608 642 836 480 636 484 468 205 ...
## $ GarageQual : Factor w/ 5 levels "Ex","Fa","Gd",...: 5 5 5 5 5 5 5 5 2 3 ...
## $ GarageCond : Factor w/ 5 levels "Ex","Fa","Gd",...: 5 5 5 5 5 5 5 5 5 5 ...
## $ PavedDrive : Factor w/ 3 levels "N","P","Y": 3 3 3 3 3 3 3 3 3 3 ...
## $ WoodDeckSF : int 0 298 0 0 192 40 255 235 90 0 ...
## $ OpenPorchSF : int 61 0 42 35 84 30 57 204 0 4 ...
## $ EnclosedPorch : int 0 0 0 272 0 0 0 228 205 0 ...
## $ X3SsnPorch : int 0 0 0 0 0 320 0 0 0 0 ...
## $ ScreenPorch : int 0 0 0 0 0 0 0 0 0 0 ...
## $ PoolArea : int 0 0 0 0 0 0 0 0 0 0 ...
## $ PoolQC : Factor w/ 3 levels "Ex","Fa","Gd": NA NA NA NA NA NA NA NA NA NA ...
## $ Fence : Factor w/ 4 levels "GdPrv","GdWo",...: NA NA NA NA NA 3 NA NA NA NA ...
## $ MiscFeature : Factor w/ 4 levels "Gar2","Othr",...: NA NA NA NA NA 3 NA 3 NA NA ...
## $ MiscVal : int 0 0 0 0 0 700 0 350 0 0 ...
## $ MoSold : int 2 5 9 2 12 10 8 11 4 1 ...

```

```
## $ YrSold      : int  2008 2007 2008 2006 2008 2009 2007 2009 2008 2008 ...
## $ SaleType    : Factor w/ 9 levels "COD","Con","ConLD",...: 9 9 9 9 9 9 9 9 9 ...
## $ SaleCondition: Factor w/ 6 levels "Abnorml","AdjLand",...: 5 5 5 1 5 5 5 1 5 ...
## $ SalePrice   : int  208500 181500 223500 140000 250000 143000 307000 200000 129900 118000 ...
```

```
summary(train_data)
```

```
##      Id      MSSubClass      MSZoning      LotFrontage
## Min.   : 1.0   Min.   : 20.0   C (all): 10   Min.   : 21.00
## 1st Qu.: 365.8 1st Qu.: 20.0   FV      : 65   1st Qu.: 59.00
## Median : 730.5 Median : 50.0   RH      : 16   Median : 69.00
## Mean   : 730.5 Mean   : 56.9   RL      :1151   Mean   : 70.05
## 3rd Qu.:1095.2 3rd Qu.: 70.0   RM      : 218   3rd Qu.: 80.00
## Max.   :1460.0 Max.   :190.0           Max.   :313.00
##                                     NA's   :259
##      LotArea      Street      Alley      LotShape LandContour Utilities
## Min.   : 1300   Grvl: 6   Grvl: 50   IR1:484   Bnk: 63   AllPub:1459
## 1st Qu.: 7554   Pave:1454 Pave: 41   IR2: 41   HLS: 50   NoSeWa: 1
## Median : 9478           NA's:1369   IR3: 10   Low: 36
## Mean   : 10517           Reg:925   Lvl:1311
## 3rd Qu.: 11602
## Max.   :215245
##
##      LotConfig      LandSlope      Neighborhood      Condition1      Condition2
## Corner : 263   Gtl:1382   NAmes :225   Norm :1260   Norm :1445
## CulDSac: 94   Mod: 65   CollgCr:150   Feedr : 81   Feedr : 6
## FR2 : 47   Sev: 13   OldTown:113   Artery : 48   Artery : 2
## FR3 : 4           Edwards:100   RRAn : 26   PosN : 2
## Inside :1052   Somerst: 86   PosN : 19   RRNn : 2
##           Gilbert: 79   RRAe : 11   PosA : 1
##           (Other):707   (Other): 15   (Other): 2
##      BldgType      HouseStyle      OverallQual      OverallCond      YearBuilt
## 1Fam :1220   1Story :726   Min. : 1.000   Min. :1.000   Min. :1872
## 2fmCon: 31   2Story :445   1st Qu.: 5.000   1st Qu.:5.000   1st Qu.:1954
## Duplex: 52   1.5Fin :154   Median : 6.000   Median :5.000   Median :1973
## Twnhs : 43   SLvl : 65   Mean : 6.099   Mean :5.575   Mean :1971
## TwnhsE: 114   SFoyer : 37   3rd Qu.: 7.000   3rd Qu.:6.000   3rd Qu.:2000
##           1.5Unf : 14   Max. :10.000   Max. :9.000   Max. :2010
##           (Other): 19
##      YearRemodAdd      RoofStyle      RoofMatl      Exterior1st      Exterior2nd
## Min. :1950   Flat : 13   CompShg:1434   VinylSd:515   VinylSd:504
## 1st Qu.:1967   Gable :1141   Tar&Grv: 11   HdBoard:222   MetalSd:214
## Median :1994   Gambrel: 11   WdShngl: 6   MetalSd:220   HdBoard:207
## Mean :1985   Hip : 286   WdShake: 5   Wd Sdng:206   Wd Sdng:197
## 3rd Qu.:2004   Mansard: 7   ClyTile: 1   Plywood:108   Plywood:142
## Max. :2010   Shed : 2   Membran: 1   CemntBd: 61   CmentBd: 60
##           (Other): 2   (Other):128   (Other):136
##      MasVnrType      MasVnrArea      ExterQual      ExterCond      Foundation      BsmtQual
## BrkCmn : 15   Min. : 0.0   Ex: 52   Ex: 3   BrkTil:146   Ex :121
## BrkFace:445   1st Qu.: 0.0   Fa: 14   Fa: 28   CBlock:634   Fa : 35
## None :864   Median : 0.0   Gd:488   Gd: 146   PConc :647   Gd :618
## Stone :128   Mean : 103.7   TA:906   Po: 1   Slab : 24   TA :649
## NA's : 8   3rd Qu.: 166.0           TA:1282   Stone : 6   NA's: 37
##           Max. :1600.0           Wood : 3
##           NA's :8
```

```

## BsmtCond      BsmtExposure BsmtFinType1      BsmtFinSF1      BsmtFinType2
## Fa : 45      Av :221      ALQ :220      Min. : 0.0      ALQ : 19
## Gd : 65      Gd :134      BLQ :148      1st Qu.: 0.0      BLQ : 33
## Po : 2       Mn :114      GLQ :418      Median : 383.5      GLQ : 14
## TA :1311     No :953      LwQ : 74      Mean : 443.6      LwQ : 46
## NA's: 37     NA's: 38      Rec :133      3rd Qu.: 712.2      Rec : 54
##                                     Unf :430      Max. :5644.0      Unf :1256
##                                     NA's: 37      NA's: 38
## BsmtFinSF2      BsmtUnfSF      TotalBsmtSF      Heating      HeatingQC
## Min. : 0.00      Min. : 0.0      Min. : 0.0      Floor: 1      Ex:741
## 1st Qu.: 0.00      1st Qu.: 223.0      1st Qu.: 795.8      GasA :1428      Fa: 49
## Median : 0.00      Median : 477.5      Median : 991.5      GasW : 18      Gd:241
## Mean : 46.55      Mean : 567.2      Mean :1057.4      Grav : 7      Po: 1
## 3rd Qu.: 0.00      3rd Qu.: 808.0      3rd Qu.:1298.2      OthW : 2      TA:428
## Max. :1474.00      Max. :2336.0      Max. :6110.0      Wall : 4
##
## CentralAir Electrical      X1stFlrSF      X2ndFlrSF      LowQualFinSF
## N: 95      FuseA: 94      Min. : 334      Min. : 0      Min. : 0.000
## Y:1365      FuseF: 27      1st Qu.: 882      1st Qu.: 0      1st Qu.: 0.000
##                                     FuseP: 3      Median :1087      Median : 0      Median : 0.000
##                                     Mix : 1      Mean :1163      Mean : 347      Mean : 5.845
##                                     SBrkr:1334      3rd Qu.:1391      3rd Qu.: 728      3rd Qu.: 0.000
##                                     NA's : 1      Max. :4692      Max. :2065      Max. :572.000
##
## GrLivArea      BsmtFullBath      BsmtHalfBath      FullBath
## Min. : 334      Min. :0.0000      Min. :0.00000      Min. :0.000
## 1st Qu.:1130      1st Qu.:0.0000      1st Qu.:0.00000      1st Qu.:1.000
## Median :1464      Median :0.0000      Median :0.00000      Median :2.000
## Mean :1515      Mean :0.4253      Mean :0.05753      Mean :1.565
## 3rd Qu.:1777      3rd Qu.:1.0000      3rd Qu.:0.00000      3rd Qu.:2.000
## Max. :5642      Max. :3.0000      Max. :2.00000      Max. :3.000
##
## HalfBath      BedroomAbvGr      KitchenAbvGr      KitchenQual      TotRmsAbvGrd
## Min. :0.0000      Min. :0.000      Min. :0.000      Ex:100      Min. : 2.000
## 1st Qu.:0.0000      1st Qu.:2.000      1st Qu.:1.000      Fa: 39      1st Qu.: 5.000
## Median :0.0000      Median :3.000      Median :1.000      Gd:586      Median : 6.000
## Mean :0.3829      Mean :2.866      Mean :1.047      TA:735      Mean : 6.518
## 3rd Qu.:1.0000      3rd Qu.:3.000      3rd Qu.:1.000      3rd Qu.: 7.000
## Max. :2.0000      Max. :8.000      Max. :3.000      Max. :14.000
##
## Functional      Fireplaces      FireplaceQu      GarageType      GarageYrBlt
## Maj1: 14      Min. :0.000      Ex : 24      2Types : 6      Min. :1900
## Maj2: 5      1st Qu.:0.000      Fa : 33      Attchd :870      1st Qu.:1961
## Min1: 31      Median :1.000      Gd :380      Basement: 19      Median :1980
## Min2: 34      Mean :0.613      Po : 20      BuiltIn: 88      Mean :1979
## Mod : 15      3rd Qu.:1.000      TA :313      CarPort: 9      3rd Qu.:2002
## Sev : 1      Max. :3.000      NA's:690      Detchd :387      Max. :2010
## Typ :1360      NA's : 81      NA's :81
## GarageFinish      GarageCars      GarageArea      GarageQual      GarageCond
## Fin :352      Min. :0.000      Min. : 0.0      Ex : 3      Ex : 2
## RFn :422      1st Qu.:1.000      1st Qu.: 334.5      Fa : 48      Fa : 35
## Unf :605      Median :2.000      Median : 480.0      Gd : 14      Gd : 9
## NA's: 81      Mean :1.767      Mean : 473.0      Po : 3      Po : 7
##                                     3rd Qu.:2.000      3rd Qu.: 576.0      TA :1311      TA :1326

```

```

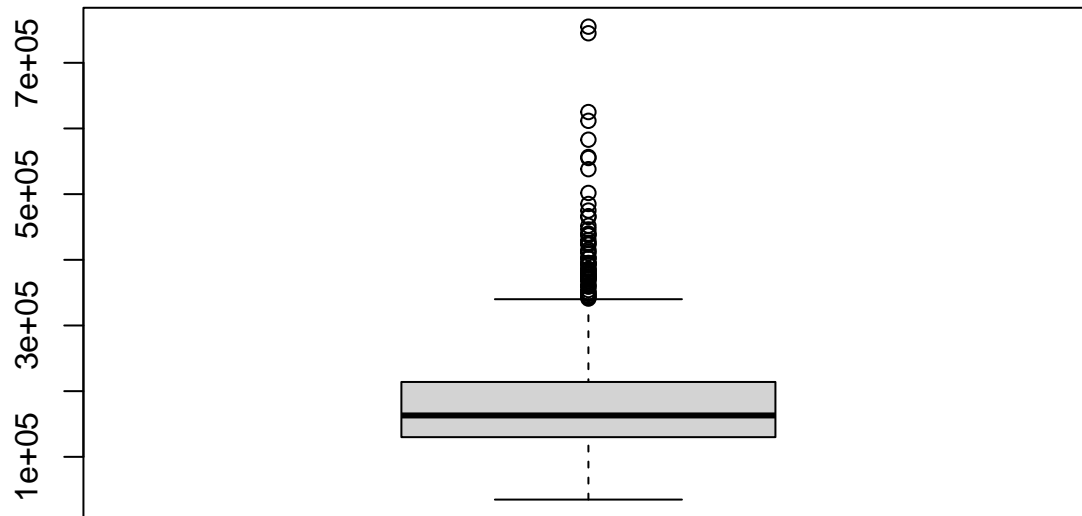
##           Max.    :4.000   Max.    :1418.0   NA's: 81   NA's: 81
##
## PavedDrive   WoodDeckSF      OpenPorchSF      EnclosedPorch      X3SsnPorch
## N: 90      Min.    : 0.00   Min.    : 0.00   Min.    : 0.00   Min.    : 0.00
## P: 30      1st Qu.: 0.00   1st Qu.: 0.00   1st Qu.: 0.00   1st Qu.: 0.00
## Y:1340     Median : 0.00   Median : 25.00   Median : 0.00   Median : 0.00
##           Mean    : 94.24   Mean    : 46.66   Mean    : 21.95   Mean    : 3.41
##           3rd Qu.:168.00   3rd Qu.: 68.00   3rd Qu.: 0.00   3rd Qu.: 0.00
##           Max.    :857.00   Max.    :547.00   Max.    :552.00   Max.    :508.00
##
## ScreenPorch   PoolArea      PoolQC      Fence      MiscFeature
## Min.    : 0.00   Min.    : 0.000   Ex   : 2   GdPrv: 59   Gar2: 2
## 1st Qu.: 0.00   1st Qu.: 0.000   Fa   : 2   GdWo : 54   Othr: 2
## Median : 0.00   Median : 0.000   Gd   : 3   MnPrv: 157   Shed: 49
## Mean    : 15.06   Mean    : 2.759   NA's:1453   MnWw : 11   TenC: 1
## 3rd Qu.: 0.00   3rd Qu.: 0.000   NA's :1179   NA's:1406
## Max.    :480.00   Max.    :738.000
##
## MiscVal      MoSold      YrSold      SaleType
## Min.    : 0.00   Min.    : 1.000   Min.    :2006   WD      :1267
## 1st Qu.: 0.00   1st Qu.: 5.000   1st Qu.:2007   New     : 122
## Median : 0.00   Median : 6.000   Median :2008   COD     : 43
## Mean    : 43.49   Mean    : 6.322   Mean    :2008   ConLD   : 9
## 3rd Qu.: 0.00   3rd Qu.: 8.000   3rd Qu.:2009   ConLI   : 5
## Max.    :15500.00   Max.    :12.000   Max.    :2010   ConLw   : 5
##                                     (Other): 9
## SaleCondition   SalePrice
## Abnorml: 101   Min.    : 34900
## AdjLand: 4    1st Qu.:129975
## Alloca : 12   Median :163000
## Family : 20   Mean    :180921
## Normal :1198   3rd Qu.:214000
## Partial: 125   Max.    :755000
##

```

Visualizing the Data

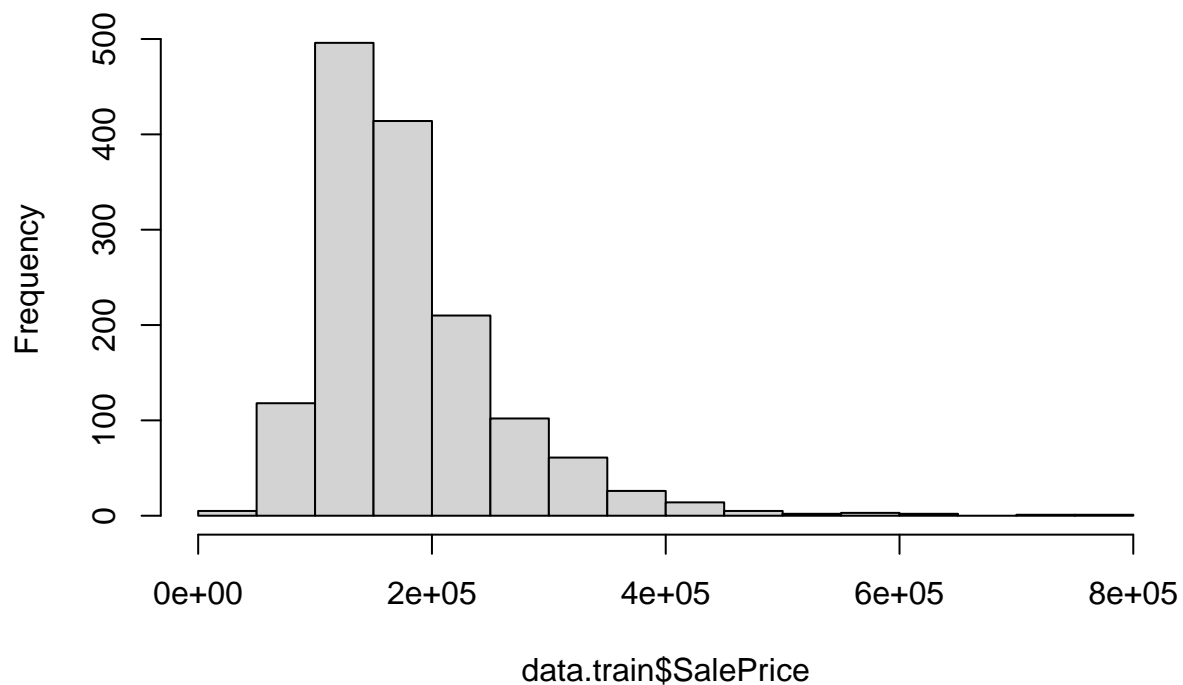
To check the distribution of the target variable `SalePrice`, we will look at a boxplot and a histogram as shown below:

```
boxplot(data.train$SalePrice)
```



```
hist(data.train$SalePrice)
```

Histogram of data.train\$SalePrice



Correlation with SalePrice

To analyze the data, we will need to create a model using the `lm()` function, where we will be using `SalePrice` to compare our other fields. Using the model, we can use `summary()` to find better information about our model, such as R^2 .

```
model <- lm(SalePrice ~ ., data=data.train)
summary(model)
```

```
##
## Call:
## lm(formula = SalePrice ~ ., data = data.train)
```

```

##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -174481  -10672       87    9739  174481
##
## Coefficients: (3 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.177e+06  1.062e+06  -1.108  0.267967
## MSSubClass    -9.706e+00  8.521e+01  -0.114  0.909331
## MSZoningFV     3.090e+04  1.218e+04   2.537  0.011307 *
## MSZoningRH     2.403e+04  1.226e+04   1.961  0.050124 .
## MSZoningRL     2.591e+04  1.044e+04   2.482  0.013187 *
## MSZoningRM     2.498e+04  9.768e+03   2.558  0.010652 *
## LotArea        7.020e-01  1.078e-01   6.511  1.07e-10 ***
## StreetPave     3.886e+04  1.225e+04   3.172  0.001552 **
## LotShapeIR2     4.488e+03  4.310e+03   1.041  0.298026
## LotShapeIR3     3.606e+03  8.782e+03   0.411  0.681391
## LotShapeReg     5.814e+02  1.660e+03   0.350  0.726120
## LandContourHLS  1.377e+04  5.274e+03   2.611  0.009138 **
## LandContourLow -4.102e+03  6.505e+03  -0.631  0.528439
## LandContourLvl  7.264e+03  3.799e+03   1.912  0.056096 .
## UtilitiesNoSeWa -2.945e+04  2.638e+04  -1.117  0.264409
## LotConfigCulDSac 7.684e+03  3.317e+03   2.316  0.020706 *
## LotConfigFR2    -5.783e+03  4.146e+03  -1.395  0.163346
## LotConfigFR3    -1.338e+04  1.305e+04  -1.026  0.305262
## LotConfigInside -1.223e+03  1.800e+03  -0.680  0.496724
## LandSlopeMod     1.054e+04  4.025e+03   2.618  0.008960 **
## LandSlopeSev    -2.553e+04  1.106e+04  -2.308  0.021169 *
## NeighborhoodBlueste -2.591e+03  1.931e+04  -0.134  0.893306
## NeighborhoodBrDale  8.307e+03  1.110e+04   0.748  0.454555
## NeighborhoodBrkSide -1.848e+03  9.474e+03  -0.195  0.845418
## NeighborhoodClearCr -1.285e+04  9.402e+03  -1.366  0.172110
## NeighborhoodCollgCr -9.675e+03  7.320e+03  -1.322  0.186510
## NeighborhoodCrawfor 9.576e+03  8.653e+03   1.107  0.268642
## NeighborhoodEdwards -1.675e+04  8.063e+03  -2.078  0.037942 *
## NeighborhoodGilbert -1.372e+04  7.831e+03  -1.752  0.079967 .
## NeighborhoodIDOTRR -7.378e+03  1.082e+04  -0.682  0.495261
## NeighborhoodMeadowV -1.610e+03  1.138e+04  -0.141  0.887530
## NeighborhoodMitchel -2.036e+04  8.262e+03  -2.464  0.013853 *
## NeighborhoodNames -1.449e+04  7.887e+03  -1.837  0.066426 .
## NeighborhoodNoRidge 2.889e+04  8.381e+03   3.447  0.000585 ***
## NeighborhoodNPkVill 8.215e+03  1.430e+04   0.574  0.565891
## NeighborhoodNridgHt 2.457e+04  7.363e+03   3.337  0.000871 ***
## NeighborhoodNWAmes -2.050e+04  8.129e+03  -2.522  0.011802 *
## NeighborhoodOldTown -1.300e+04  9.654e+03  -1.346  0.178498
## NeighborhoodSawyer -1.020e+04  8.214e+03  -1.241  0.214694
## NeighborhoodSawyerW -6.161e+03  7.840e+03  -0.786  0.432106
## NeighborhoodSomerst 1.701e+02  8.956e+03   0.019  0.984850
## NeighborhoodStoneBr 3.894e+04  8.369e+03   4.652  3.63e-06 ***
## NeighborhoodSWISU -9.499e+03  9.808e+03  -0.968  0.333002
## NeighborhoodTimber -5.900e+03  8.352e+03  -0.706  0.480067
## NeighborhoodVeenker 3.057e+03  1.071e+04   0.285  0.775465
## Condition1Feedr  3.035e+03  5.088e+03   0.597  0.550888
## Condition1Norm    1.221e+04  4.194e+03   2.912  0.003652 **

```

## Condition1PosA	7.473e+03	1.028e+04	0.727	0.467294	
## Condition1PosN	8.077e+03	7.598e+03	1.063	0.287986	
## Condition1RR Ae	-1.693e+04	9.355e+03	-1.810	0.070591	.
## Condition1RR An	6.490e+03	6.998e+03	0.927	0.353893	
## Condition1RR Ne	-7.278e+03	1.834e+04	-0.397	0.691557	
## Condition1RR Nn	3.894e+03	1.308e+04	0.298	0.765911	
## Condition2Feedr	-9.591e+03	2.300e+04	-0.417	0.676776	
## Condition2Norm	-7.506e+03	1.958e+04	-0.383	0.701533	
## Condition2PosA	1.981e+04	3.793e+04	0.522	0.601677	
## Condition2PosN	-2.303e+05	2.756e+04	-8.355	< 2e-16	***
## Condition2RR Ae	-1.290e+05	4.672e+04	-2.762	0.005826	**
## Condition2RR An	-1.203e+04	3.188e+04	-0.377	0.706040	
## Condition2RR Nn	-9.030e+03	2.705e+04	-0.334	0.738563	
## BldgType2fmCon	-6.227e+03	1.285e+04	-0.485	0.627914	
## BldgTypeDuplex	-1.012e+03	7.434e+03	-0.136	0.891695	
## BldgTypeTwnhs	-2.535e+04	1.013e+04	-2.503	0.012452	*
## BldgTypeTwnhsE	-2.313e+04	9.172e+03	-2.522	0.011800	*
## HouseStyle1.5Unf	1.091e+04	7.882e+03	1.384	0.166513	
## HouseStyle1Story	8.847e+03	4.321e+03	2.047	0.040846	*
## HouseStyle2.5Fin	-1.702e+04	1.227e+04	-1.387	0.165727	
## HouseStyle2.5Unf	-1.154e+04	9.362e+03	-1.233	0.217944	
## HouseStyle2Story	-6.393e+03	3.541e+03	-1.805	0.071281	.
## HouseStyleSFoyer	7.780e+03	6.178e+03	1.259	0.208155	
## HouseStyleSLvl	7.395e+03	5.447e+03	1.358	0.174759	
## OverallQual	7.995e+03	1.016e+03	7.866	7.75e-15	***
## OverallCond	5.378e+03	8.690e+02	6.189	8.15e-10	***
## YearBuilt	3.255e+02	7.342e+01	4.433	1.01e-05	***
## YearRemodAdd	1.048e+02	5.520e+01	1.899	0.057764	.
## RoofStyleGable	1.166e+03	1.871e+04	0.062	0.950327	
## RoofStyleGambrel	4.014e+03	2.047e+04	0.196	0.844550	
## RoofStyleHip	2.794e+03	1.876e+04	0.149	0.881604	
## RoofStyleMansard	1.701e+04	2.180e+04	0.780	0.435486	
## RoofStyleShed	8.806e+04	3.544e+04	2.485	0.013079	*
## RoofMatlCompShg	6.490e+05	3.290e+04	19.726	< 2e-16	***
## RoofMatlMembran	7.357e+05	4.761e+04	15.453	< 2e-16	***
## RoofMatlMetal	6.956e+05	4.702e+04	14.793	< 2e-16	***
## RoofMatlRoll	6.501e+05	4.149e+04	15.670	< 2e-16	***
## RoofMatlTar&Grv	6.542e+05	3.782e+04	17.297	< 2e-16	***
## RoofMatlWdShake	6.290e+05	3.662e+04	17.178	< 2e-16	***
## RoofMatlWdShngl	7.264e+05	3.414e+04	21.279	< 2e-16	***
## Exterior1stAsphShn	-1.052e+04	3.383e+04	-0.311	0.755948	
## Exterior1stBrkComm	-1.114e+04	2.831e+04	-0.394	0.693978	
## Exterior1stBrkFace	6.826e+03	1.251e+04	0.546	0.585378	
## Exterior1stCBlock	-2.771e+04	2.751e+04	-1.007	0.314067	
## Exterior1stCemntBd	-1.365e+04	1.916e+04	-0.712	0.476430	
## Exterior1stHdBoard	-1.240e+04	1.257e+04	-0.986	0.324266	
## Exterior1stImStucc	-6.737e+04	2.840e+04	-2.372	0.017821	*
## Exterior1stMetalSd	-1.769e+03	1.447e+04	-0.122	0.902709	
## Exterior1stPlywood	-1.651e+04	1.245e+04	-1.327	0.184821	
## Exterior1stStone	-1.357e+04	2.413e+04	-0.562	0.573885	
## Exterior1stStucco	-3.662e+03	1.377e+04	-0.266	0.790369	
## Exterior1stVinylSd	-1.633e+04	1.317e+04	-1.240	0.215151	
## Exterior1stWd Sdng	-1.239e+04	1.204e+04	-1.029	0.303605	
## Exterior1stWdShing	-4.981e+03	1.304e+04	-0.382	0.702524	

## Exterior2ndAsphShn	6.790e+03	2.262e+04	0.300	0.764070	
## Exterior2ndBrk Cmn	1.390e+04	2.059e+04	0.675	0.499800	
## Exterior2ndBrkFace	-1.477e+03	1.318e+04	-0.112	0.910796	
## Exterior2ndCBlock	NA	NA	NA	NA	
## Exterior2ndCmentBd	1.242e+04	1.909e+04	0.651	0.515346	
## Exterior2ndHdBoard	7.218e+03	1.233e+04	0.585	0.558422	
## Exterior2ndImStucc	3.276e+04	1.430e+04	2.290	0.022173	*
## Exterior2ndMetalSd	2.064e+03	1.430e+04	0.144	0.885259	
## Exterior2ndOther	-6.816e+03	2.812e+04	-0.242	0.808544	
## Exterior2ndPlywood	8.246e+03	1.197e+04	0.689	0.491084	
## Exterior2ndStone	-1.090e+04	1.719e+04	-0.634	0.525999	
## Exterior2ndStucco	1.788e+03	1.355e+04	0.132	0.895032	
## Exterior2ndVinylSd	1.576e+04	1.290e+04	1.222	0.221933	
## Exterior2ndWd Sdng	9.923e+03	1.183e+04	0.839	0.401806	
## Exterior2ndWd Shng	2.630e+03	1.236e+04	0.213	0.831584	
## ExterQualFa	-8.529e+03	1.084e+04	-0.787	0.431681	
## ExterQualGd	-3.090e+04	4.781e+03	-6.462	1.47e-10	***
## ExterQualTA	-3.084e+04	5.350e+03	-5.765	1.03e-08	***
## ExterCondFa	-2.928e+03	1.881e+04	-0.156	0.876302	
## ExterCondGd	-8.101e+03	1.798e+04	-0.451	0.652382	
## ExterCondPo	1.141e+04	3.272e+04	0.349	0.727454	
## ExterCondTA	-5.396e+03	1.795e+04	-0.301	0.763722	
## FoundationCBlock	1.815e+03	3.181e+03	0.571	0.568281	
## FoundationPConc	4.844e+03	3.497e+03	1.385	0.166262	
## FoundationSlab	8.598e+03	7.831e+03	1.098	0.272467	
## FoundationStone	1.122e+03	1.086e+04	0.103	0.917690	
## FoundationWood	-3.322e+04	1.509e+04	-2.202	0.027845	*
## BsmtFinSF1	3.705e+01	4.406e+00	8.410	< 2e-16	***
## BsmtFinSF2	2.450e+01	5.784e+00	4.235	2.44e-05	***
## BsmtUnfSF	1.494e+01	4.058e+00	3.682	0.000242	***
## TotalBsmtSF	NA	NA	NA	NA	
## HeatingGasA	-6.917e+03	2.535e+04	-0.273	0.784959	
## HeatingGasW	-1.549e+04	2.616e+04	-0.592	0.553999	
## HeatingGrav	-1.447e+04	2.746e+04	-0.527	0.598290	
## HeatingOthW	-4.535e+04	3.165e+04	-1.433	0.152168	
## HeatingWall	8.977e+03	2.920e+04	0.307	0.758549	
## HeatingQCFA	-1.475e+03	4.796e+03	-0.307	0.758543	
## HeatingQCGd	-3.698e+03	2.135e+03	-1.732	0.083586	.
## HeatingQCPo	7.969e+03	2.742e+04	0.291	0.771402	
## HeatingQCTA	-4.347e+03	2.116e+03	-2.055	0.040104	*
## CentralAirY	-3.391e+03	3.863e+03	-0.878	0.380291	
## X1stFlrSF	5.514e+01	5.316e+00	10.371	< 2e-16	***
## X2ndFlrSF	7.004e+01	5.237e+00	13.374	< 2e-16	***
## LowQualFinSF	2.453e+01	1.857e+01	1.321	0.186673	
## GrLivArea	NA	NA	NA	NA	
## BsmtFullBath	1.544e+03	1.964e+03	0.787	0.431715	
## BsmtHalfBath	3.704e+02	3.108e+03	0.119	0.905148	
## FullBath	2.525e+03	2.236e+03	1.129	0.258993	
## HalfBath	-1.492e+02	2.131e+03	-0.070	0.944200	
## BedroomAbvGr	-5.493e+03	1.378e+03	-3.987	7.06e-05	***
## KitchenAbvGr	-1.589e+04	5.717e+03	-2.780	0.005517	**
## KitchenQualFa	-2.110e+04	6.340e+03	-3.329	0.000898	***
## KitchenQualGd	-2.780e+04	3.480e+03	-7.987	3.06e-15	***
## KitchenQualTA	-2.536e+04	3.984e+03	-6.366	2.70e-10	***

```
## TotRmsAbvGrd      1.366e+03  9.726e+02  1.405 0.160408
## FunctionalMaj2    2.301e+02  1.359e+04  0.017 0.986491
## FunctionalMin1    4.184e+03  8.613e+03  0.486 0.627206
## FunctionalMin2    8.542e+03  8.555e+03  0.998 0.318234
## FunctionalMod     -7.163e+03  1.054e+04 -0.679 0.496974
## FunctionalSev     -6.026e+04  2.752e+04 -2.190 0.028719 *
## FunctionalTyp     1.963e+04  7.387e+03  2.657 0.007981 **
## Fireplaces        2.758e+03  1.369e+03  2.015 0.044125 *
## GarageCars        4.296e+03  2.216e+03  1.939 0.052730 .
## GarageArea        1.313e+01  7.612e+00  1.724 0.084874 .
## PavedDriveP       -3.247e+03  5.558e+03 -0.584 0.559139
## PavedDriveY       -2.036e+03  3.435e+03 -0.593 0.553525
## WoodDeckSF        1.361e+01  5.943e+00  2.290 0.022210 *
## OpenPorchSF       1.223e+01  1.180e+01  1.037 0.299976
## EnclosedPorch     5.377e+00  1.277e+01  0.421 0.673727
## X3SsnPorch        2.400e+01  2.308e+01  1.040 0.298539
## ScreenPorch       3.702e+01  1.257e+01  2.946 0.003275 **
## PoolArea          7.139e+01  1.828e+01  3.906 9.89e-05 ***
## MiscVal           -3.282e-01  1.466e+00 -0.224 0.822852
## MoSold            -6.294e+02  2.529e+02 -2.488 0.012966 *
## YrSold            -1.797e+02  5.234e+02 -0.343 0.731396
## SaleTypeCon       3.533e+04  1.835e+04  1.926 0.054373 .
## SaleTypeConLD     1.682e+04  9.993e+03  1.683 0.092642 .
## SaleTypeConLI     9.657e+03  1.188e+04  0.813 0.416551
## SaleTypeConLw     -2.273e+03  1.238e+04 -0.184 0.854415
## SaleTypeCWD       2.325e+04  1.332e+04  1.745 0.081277 .
## SaleTypeNew       3.472e+04  1.601e+04  2.169 0.030294 *
## SaleTypeOth       1.864e+04  1.498e+04  1.244 0.213637
## SaleTypeWD        5.420e+02  4.328e+03  0.125 0.900362
## SaleConditionAdjLand 8.245e+03  1.446e+04  0.570 0.568754
## SaleConditionAlloca 4.978e+03  8.751e+03  0.569 0.569512
## SaleConditionFamily -1.417e+03  6.306e+03 -0.225 0.822241
## SaleConditionNormal 6.527e+03  2.970e+03  2.198 0.028155 *
## SaleConditionPartial -9.505e+03  1.543e+04 -0.616 0.537950
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23960 on 1274 degrees of freedom
## Multiple R-squared:  0.9206, Adjusted R-squared:  0.909
## F-statistic: 79.79 on 185 and 1274 DF,  p-value: < 2.2e-16
```

With the summary we can see that we have $R^2 = 0.9206$, and we can make some of the following interpretations when comparing SalesPrice with some of the most significant fields marked with ***: Whenever the sales price increases by one dollar...

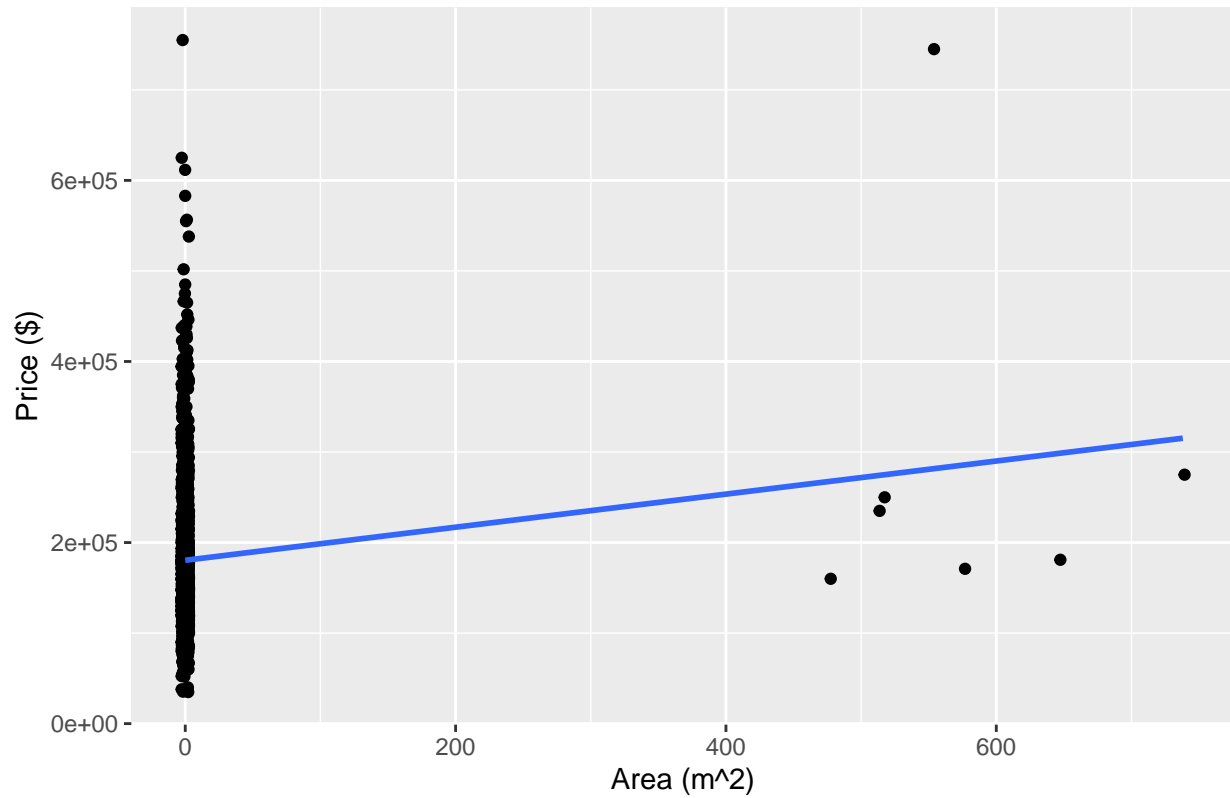
- PoolArea: the pool's area will increase by $71.39m^2$.
- OverallQual: the overall quality will increase by \$7995.
- OverallCond: the overall condition will increase by \$5378.
- LotArea: the lot's area will increase by $0.702m^2$.

For a better visual we will be plotting graphs between the sale price and the significant fields mentioned.

```
# Plot between PoolArea and SalePrice
ggplot(data = data.train, aes(x = PoolArea, y = SalePrice)) +
  geom_jitter() + geom_smooth(method = "lm", se = FALSE) +
  labs(title="Scatter plot of Pool Area and Sales Price", x="Area (m^2)", y="Price ($)")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

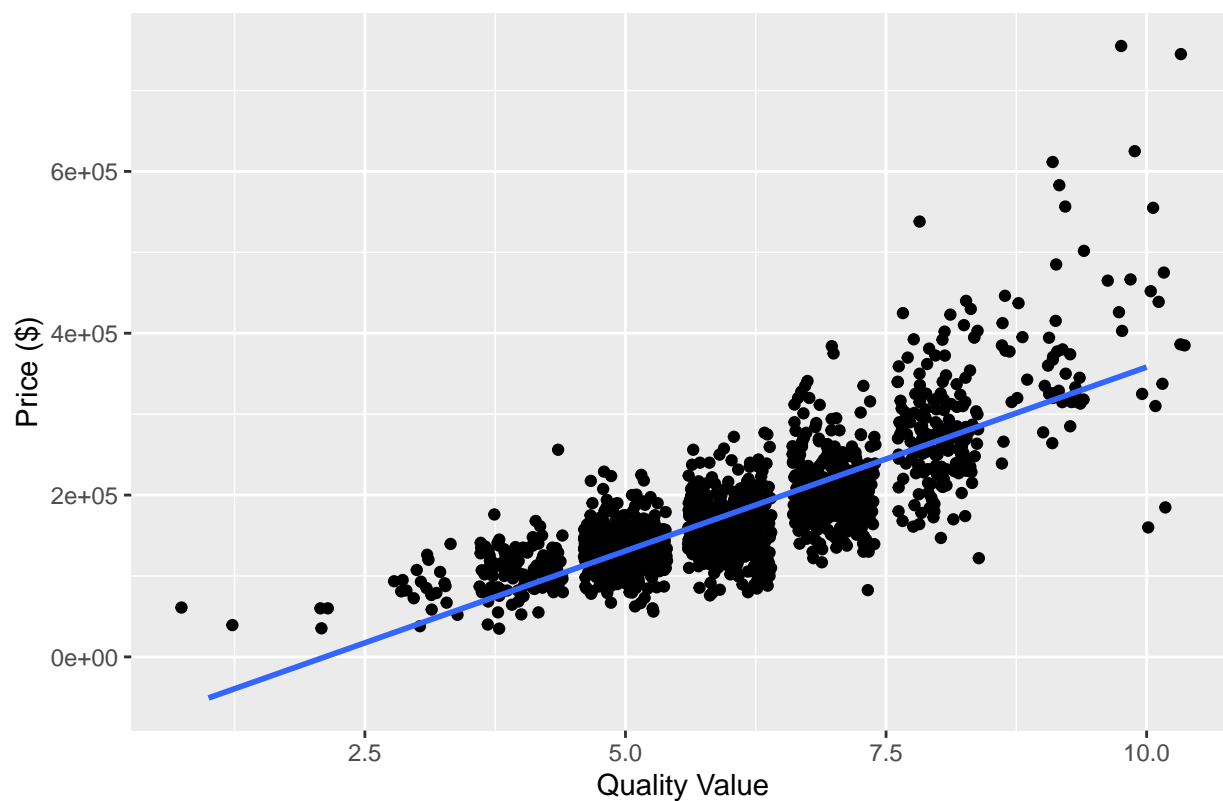
Scatter plot of Pool Area and Sales Price



```
# Plot between Overall Quality and SalePrice
ggplot(data = data.train, aes(x = OverallQual, y = SalePrice)) +
  geom_jitter() + geom_smooth(method = "lm", se = FALSE) +
  labs(title="Plot of Overall Quality and Sales Price", x="Quality Value",y="Price ($)")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

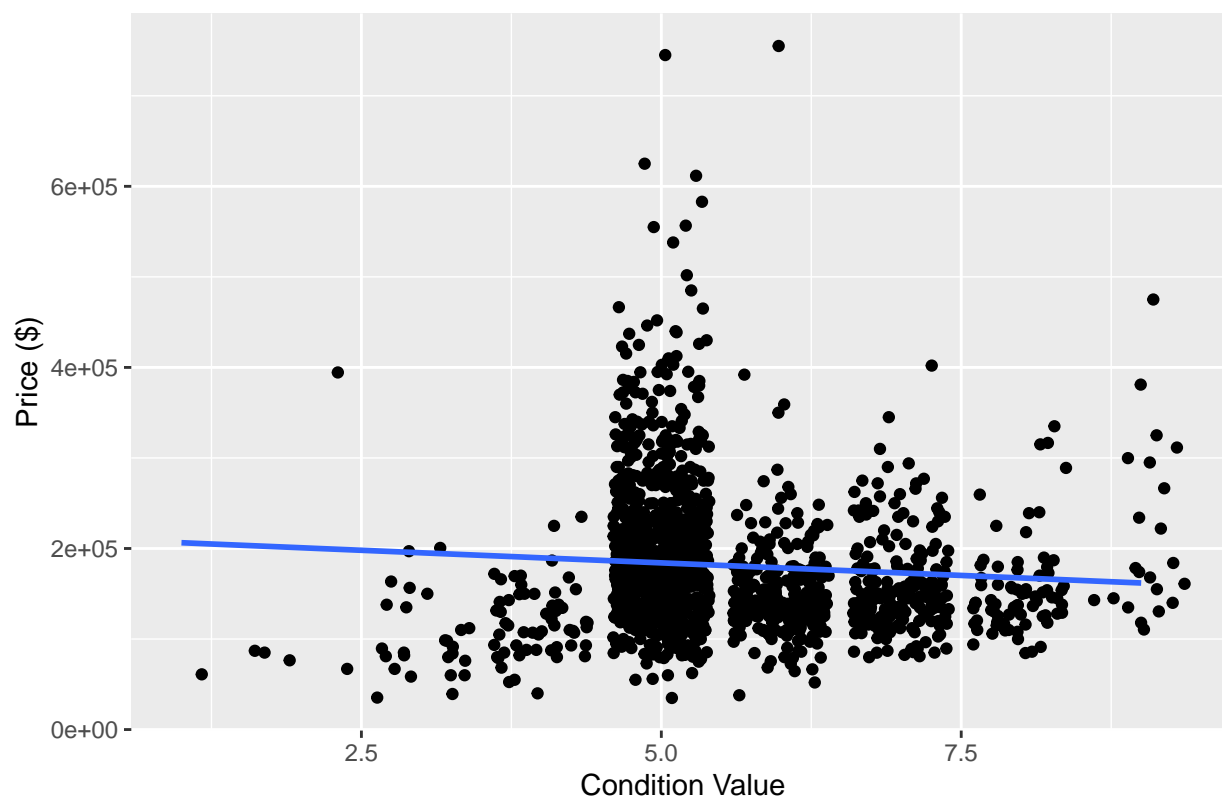
Plot of Overall Quality and Sales Price



```
# Plot between Overall Condition and SalePrice
ggplot(data = data.train, aes(x = OverallCond, y = SalePrice)) +
  geom_jitter() + geom_smooth(method = "lm", se = FALSE) +
  labs(title="Plot of Overall Condition and Sales Price", x="Condition Value",y="Price ($)")

## `geom_smooth()` using formula = 'y ~ x'
```

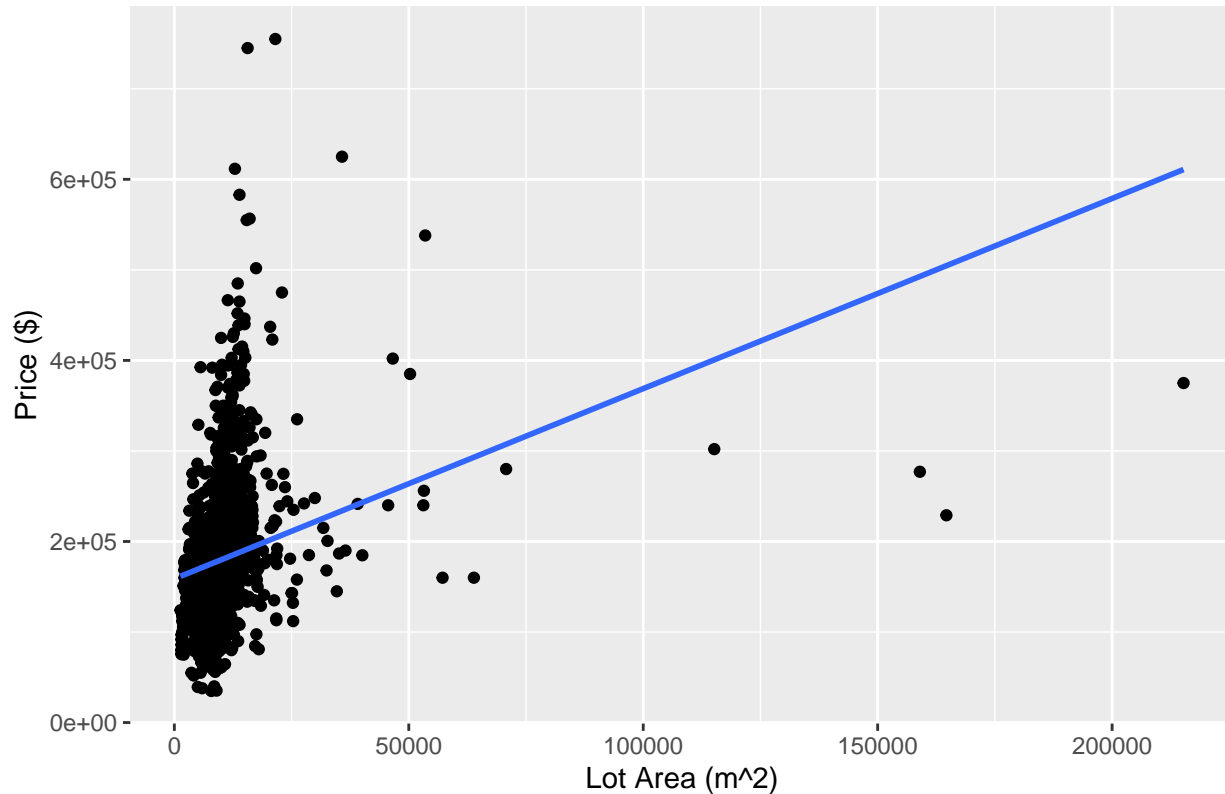
Plot of Overall Condition and Sales Price



```
# plot between Lot Area and SalePrice  
ggplot(data = data.train, aes(x = LotArea, y = SalePrice)) +  
  geom_jitter() + geom_smooth(method = "lm", se = FALSE) +  
  labs(title="Plot of Lot Area and Sales Price", x="Lot Area (m^2)", y="Price ($)")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

Plot of Lot Area and Sales Price



While our plots show a great visualization between the correlation of the Price and the mentioned significant fields, it is still a small portion of the full plate of data we have. To gain a better insight to our complete palette, we need to look into a correlation plot of all the independent numerical variables and their association with our dependent variable sales price.

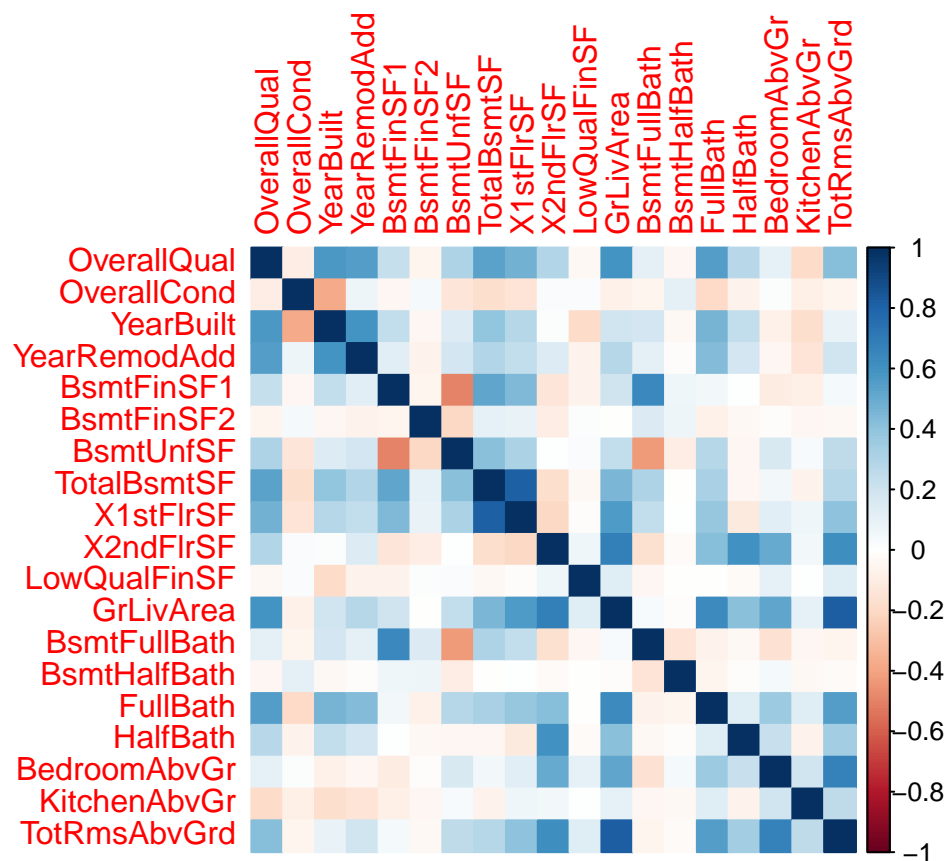
```
# get all the numerical variables
data.num <- data.train %>% dplyr::select(where(is.numeric))
# get the correlation data
data.cor <- data.frame(data.num[, 3:21])
correlation <- cor(data.cor)
# display the correlation
correlation
```

##	OverallQual	OverallCond	YearBuilt	YearRemodAdd	BsmtFinSF1
## OverallQual	1.00000000	-0.09193234	0.57232277	0.55068392	0.239665966
## OverallCond	-0.09193234	1.00000000	-0.37598320	0.07374150	-0.046230856
## YearBuilt	0.57232277	-0.37598320	1.00000000	0.59285498	0.249503197
## YearRemodAdd	0.55068392	0.07374150	0.59285498	1.00000000	0.128450547
## BsmtFinSF1	0.23966597	-0.04623086	0.24950320	0.12845055	1.000000000
## BsmtFinSF2	-0.05911869	0.04022917	-0.04910683	-0.06775851	-0.050117400
## BsmtUnfSF	0.30815893	-0.13684057	0.14904039	0.18113309	-0.495251469
## TotalBsmtSF	0.53780850	-0.17109751	0.39145200	0.29106558	0.522396052
## X1stFlrSF	0.47622383	-0.14420278	0.28198586	0.24037927	0.445862656
## X2ndFlrSF	0.29549288	0.02894212	0.01030766	0.14002378	-0.137078986
## LowQualFinSF	-0.03042928	0.02549432	-0.18378434	-0.06241910	-0.064502597
## GrLivArea	0.59300743	-0.07968587	0.19900971	0.28738852	0.208171130
## BsmtFullBath	0.11109779	-0.05494152	0.18759855	0.11946988	0.649211754
## BsmtHalfBath	-0.04015016	0.11782092	-0.03816181	-0.01233703	0.067418478

##	FullBath	0.55059971	-0.19414949	0.46827079	0.43904648	0.058543137
##	HalfBath	0.27345810	-0.06076933	0.24265591	0.18333061	0.004262424
##	BedroomAbvGr	0.10167636	0.01298006	-0.07065122	-0.04058093	-0.107354677
##	KitchenAbvGr	-0.18388223	-0.08700086	-0.17480025	-0.14959752	-0.081006851
##	TotRmsAbvGrd	0.42745234	-0.05758317	0.09558913	0.19173982	0.044315624
##		BsmtFinSF2	BsmtUnfSF	TotalBsmtSF	X1stFlrSF	X2ndFlrSF
##	OverallQual	-0.059118693	0.308158927	0.5378084986	0.476223829	0.295492879
##	OverallCond	0.040229170	-0.136840570	-0.1710975146	-0.144202784	0.028942116
##	YearBuilt	-0.049106831	0.149040392	0.3914520021	0.281985859	0.010307660
##	YearRemodAdd	-0.067758514	0.181133087	0.2910655826	0.240379268	0.140023779
##	BsmtFinSF1	-0.050117400	-0.495251469	0.5223960520	0.445862656	-0.137078986
##	BsmtFinSF2	1.000000000	-0.209294492	0.1048095376	0.097117448	-0.099260316
##	BsmtUnfSF	-0.209294492	1.000000000	0.4153596052	0.317987438	0.004469092
##	TotalBsmtSF	0.104809538	0.415359605	1.0000000000	0.819529975	-0.174511950
##	X1stFlrSF	0.097117448	0.317987438	0.8195299750	1.000000000	-0.202646181
##	X2ndFlrSF	-0.099260316	0.004469092	-0.1745119501	-0.202646181	1.000000000
##	LowQualFinSF	0.014806998	0.028166688	-0.0332453873	-0.014240673	0.063352950
##	GrLivArea	-0.009639892	0.240257268	0.4548682025	0.566023969	0.687501064
##	BsmtFullBath	0.158678061	-0.422900477	0.3073505537	0.244671104	-0.169493952
##	BsmtHalfBath	0.070948134	-0.095804288	-0.0003145818	0.001955654	-0.023854784
##	FullBath	-0.076443862	0.288886055	0.3237224136	0.380637495	0.421377983
##	HalfBath	-0.032147837	-0.041117530	-0.0488037386	-0.119915909	0.609707300
##	BedroomAbvGr	-0.015728114	0.166643317	0.0504499555	0.127400749	0.502900613
##	KitchenAbvGr	-0.040751236	0.030085868	-0.0689006426	0.068100588	0.059305753
##	TotRmsAbvGrd	-0.035226548	0.250647061	0.2855725637	0.409515979	0.616422635
##		LowQualFinSF	GrLivArea	BsmtFullBath	BsmtHalfBath	
##	OverallQual	-0.0304292840	0.593007430	0.11109779	-0.0401501577	
##	OverallCond	0.0254943199	-0.079685865	-0.05494152	0.1178209151	
##	YearBuilt	-0.1837843444	0.199009714	0.18759855	-0.0381618057	
##	YearRemodAdd	-0.0624191001	0.287388520	0.11946988	-0.0123370321	
##	BsmtFinSF1	-0.0645025969	0.208171130	0.64921175	0.0674184779	
##	BsmtFinSF2	0.0148069979	-0.009639892	0.15867806	0.0709481337	
##	BsmtUnfSF	0.0281666881	0.240257268	-0.42290048	-0.0958042882	
##	TotalBsmtSF	-0.0332453873	0.454868203	0.30735055	-0.0003145818	
##	X1stFlrSF	-0.0142406727	0.566023969	0.24467110	0.0019556536	
##	X2ndFlrSF	0.0633529501	0.687501064	-0.16949395	-0.0238547839	
##	LowQualFinSF	1.0000000000	0.134682813	-0.04714342	-0.0058415048	
##	GrLivArea	0.1346828130	1.000000000	0.03483605	-0.0189184832	
##	BsmtFullBath	-0.0471434219	0.034836050	1.00000000	-0.1478709605	
##	BsmtHalfBath	-0.0058415048	-0.018918483	-0.14787096	1.0000000000	
##	FullBath	-0.0007095096	0.630011646	-0.06451205	-0.0545358120	
##	HalfBath	-0.0270800493	0.415771636	-0.03090496	-0.0123399001	
##	BedroomAbvGr	0.1056065685	0.521269511	-0.15067281	0.0465188484	
##	KitchenAbvGr	0.0075217443	0.100063165	-0.04150255	-0.0379443502	
##	TotRmsAbvGrd	0.1311847760	0.825489374	-0.05327524	-0.0238363413	
##		FullBath	HalfBath	BedroomAbvGr	KitchenAbvGr	TotRmsAbvGrd
##	OverallQual	0.5505997094	0.273458099	0.10167636	-0.183882235	0.42745234
##	OverallCond	-0.1941494887	-0.060769327	0.01298006	-0.087000855	-0.05758317
##	YearBuilt	0.4682707872	0.242655910	-0.07065122	-0.174800246	0.09558913
##	YearRemodAdd	0.4390464839	0.183330612	-0.04058093	-0.149597521	0.19173982
##	BsmtFinSF1	0.0585431369	0.004262424	-0.10735468	-0.081006851	0.04431562
##	BsmtFinSF2	-0.0764438620	-0.032147837	-0.01572811	-0.040751236	-0.03522655
##	BsmtUnfSF	0.2888860555	-0.041117530	0.16664332	0.030085868	0.25064706
##	TotalBsmtSF	0.3237224136	-0.048803739	0.05044996	-0.068900643	0.28557256

```
## X1stFlrSF      0.3806374950 -0.119915909  0.12740075  0.068100588  0.40951598
## X2ndFlrSF      0.4213779829  0.609707300  0.50290061  0.059305753  0.61642264
## LowQualFinSF -0.0007095096 -0.027080049  0.10560657  0.007521744  0.13118478
## GrLivArea      0.6300116463  0.415771636  0.52126951  0.100063165  0.82548937
## BsmtFullBath -0.0645120486 -0.030904959 -0.15067281 -0.041502546 -0.05327524
## BsmtHalfBath -0.0545358120 -0.012339900  0.04651885 -0.037944350 -0.02383634
## FullBath      1.0000000000  0.136380589  0.36325198  0.133115214  0.55478425
## HalfBath      0.1363805887  1.0000000000  0.22665148 -0.068262549  0.34341486
## BedroomAbvGr  0.3632519830  0.226651484  1.00000000  0.198596758  0.67661994
## KitchenAbvGr  0.1331152142 -0.068262549  0.19859676  1.000000000  0.25604541
## TotRmsAbvGrd  0.5547842535  0.343414858  0.67661994  0.256045409  1.00000000
```

```
par(mfrow=c(1,1))
# plot the correlation
corrplot(correlation, method = "color")
```



Outliers

Our data is quite accurate, but to see the effects outliers may have with `SalePrice`, we will try to plot the data with and without outliers to see the effect it has on us. First we need to extract outliers from the data and obtain the data without the outliers we'll find.

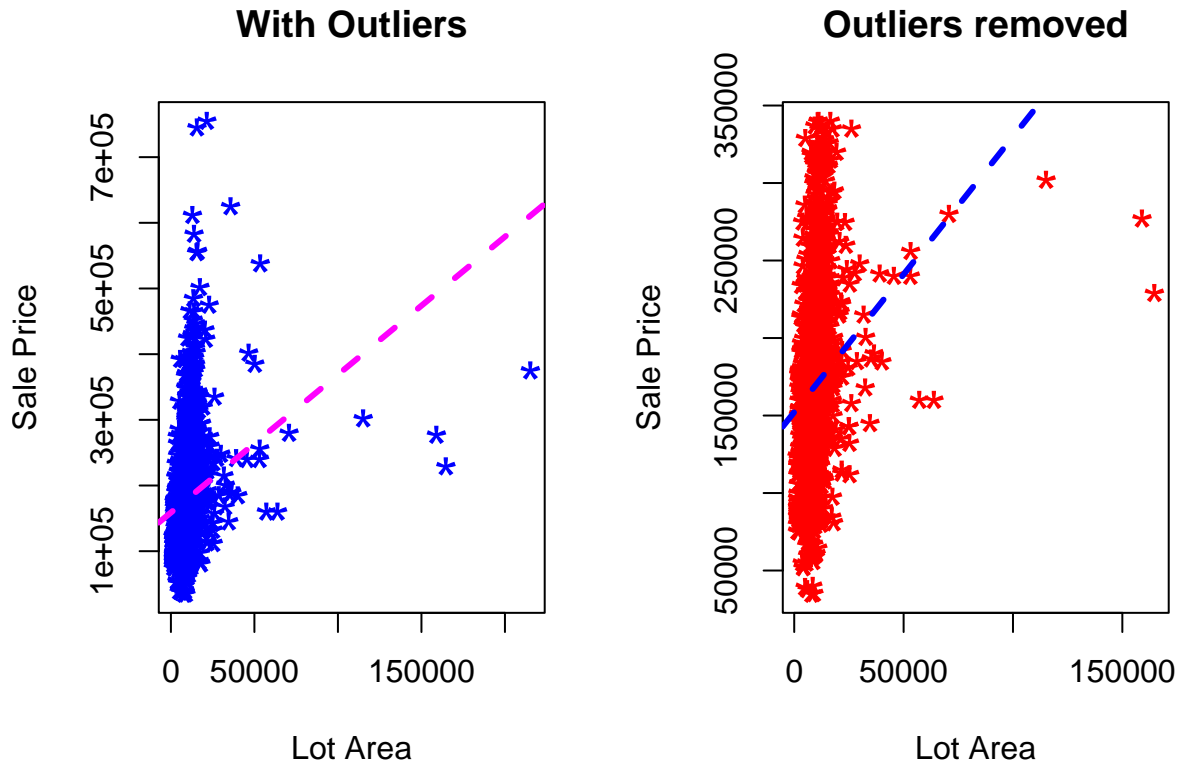
```
outliers = boxplot(data.train$SalePrice, plot = FALSE)$out
outliers_data <- data.train[which(data.train$SalePrice %in% outliers),]
no_out_train <- data.train[-which(data.train$SalePrice %in% outliers),]
```

Now we can plot our data with outliers and without to see the effect it has on our data:


```

par(mfrow=c(1,2))
# plot the data with the outliers
plot(data.train$LotArea, data.train$SalePrice, main = "With Outliers",
      xlab = "Lot Area", ylab = "Sale Price", pch="*", col = "blue", cex = 2)
abline(lm(SalePrice ~ LotArea, data=data.train), col = "magenta", lwd = 3, lty = 2)
plot(no_out_train$LotArea, no_out_train$SalePrice, main="Outliers removed",
      xlab="Lot Area", ylab="Sale Price", pch="*", col="red", cex=2)
abline(lm(SalePrice ~ LotArea, data=no_out_train), col="blue", lwd=3, lty=2)

```



As you can see, we are able to concentrate the relationship between `SalePrice` and `LotArea` by removing the outliers and find that most of the values for `SalePrice` range from around \$50000 – \$350000.

Accuracy of the Model

We will determine the accuracy of the training data model by the following:

```

# predicted values
pred <- model$fitted.values
tally_table <- data.frame(actual=data.train$SalePrice, predicted=pred)

mape <- mean(abs(tally_table$actual-tally_table$predicted)/tally_table$actual)
accuracy <- 1-mape
accuracy

```

```
## [1] 0.9151283
```

We can see that the accuracy of our training data model is 91.51%