Train dataset description

- 81 개의 속성이 포함되어 있는 데이터입니다.

- 속성들을 크게 나누면 집의 외관과 집 내관으로 나눌 수 있습니다.
- 외관: 집의 장소, 집의 종류
- 내관: 집의 크기, 방과 화장실의 갯수, 주차장의 유무, 난방과 전기

- 81 Variables (**수치형**)

MSSubClass

MSZoning

LotFrontage:

LotArea

Street

Alley

Lotshape

LandContour

Utilities

LotConfig

LandSlope

Neighborhood

Condition1

Condition2

BldgType: type of dwelling

HouseStyle: style of dwelling

OverallQual

OveerallCond

YearBuilt

YeatRemodAdd

RoofStyle

RoofMatl

Exterior1st / 2nd

MasVnrType / **Area**

ExterQual / Cond

Foundation

BsmtQual / Cond

BsmtExposure

BsmtFinType1

BsmtFinSF1

BsmtFinType2

BsmtFinSF2

BsmtUnfSF

TotalBsmtSF

Heating

HeatingQC

CentralAir

Electrical

- 81 Variables (**수치형**)

1stFlrSF

2ndFlrSF

LowQualFinSF

GrLivArea

BsmtFullBath

BsmtHalfBath

FullBath

HalfBath

BedroomAbvGr

KitchenAbvGr

KitchenQual

TotRmsAbvGrd

Functional

Fireplaces

FireplaceQu

GarageType

GarageYrBlt

GarageFinish

GarageCars

GarageArea

GarageCond

PavedDrive

WoodDeckSF

OpenPorchSF

EnclosedPorch

3SsnPorch

ScreenPorch

PoolArea

PoolQC

Fence

MiscFeature

MiscVal

MoSold

YrSold

SaleType

SaleCondition

SalePrice

Summary of Property Lot

1. LotFrontage

Min.	Q1.	Median	Mean
21.00	59.00	69.00	70.05

Q3.	Max.	NA's
80.00	313.00	259

2. LotArea

Min.	Q.1	Median	Mean
1300	7554	9478	10517

Q.3	Max.
11602	215245

Mode: 7200

3. LotShape

LotShape	Count
IR1	484
IR2	41
IR3	10
Reg	925

4. LotConfig

LotConfig	Count
Corner	263
CulDSec	94
FR2	47
FR3	4
Inside	1052

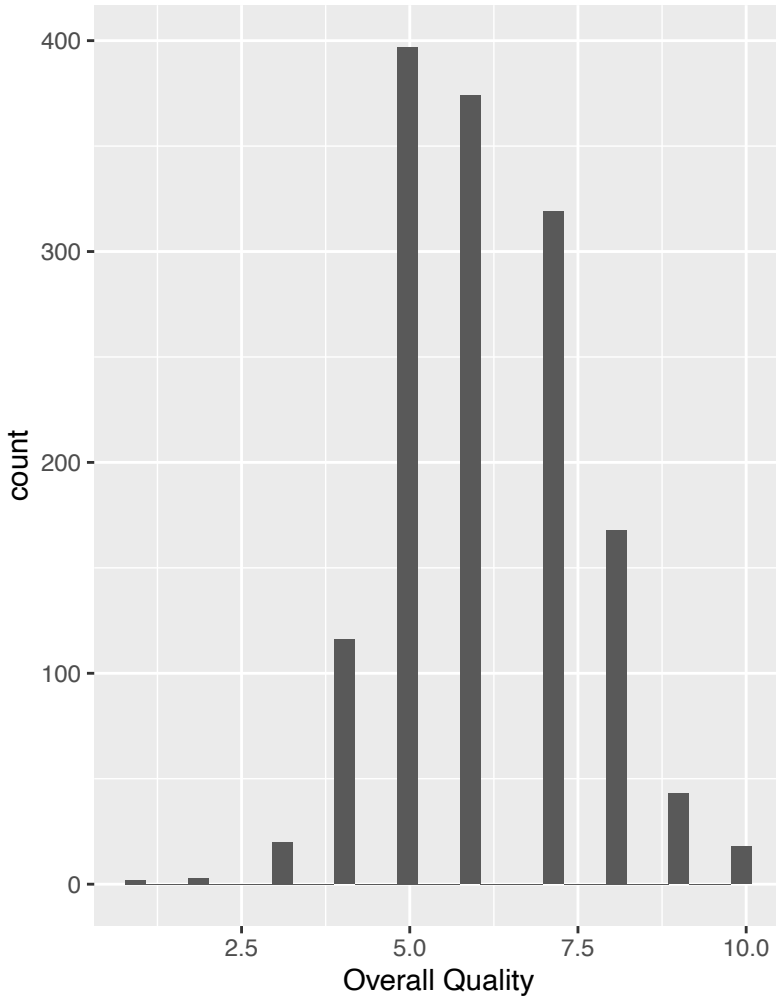
1. Type of dwelling

Type	Count
Fam	1220
2fmCon	31
Duplex	52
Twnhs	43
TwnhsE	114

2. Style of dwelling

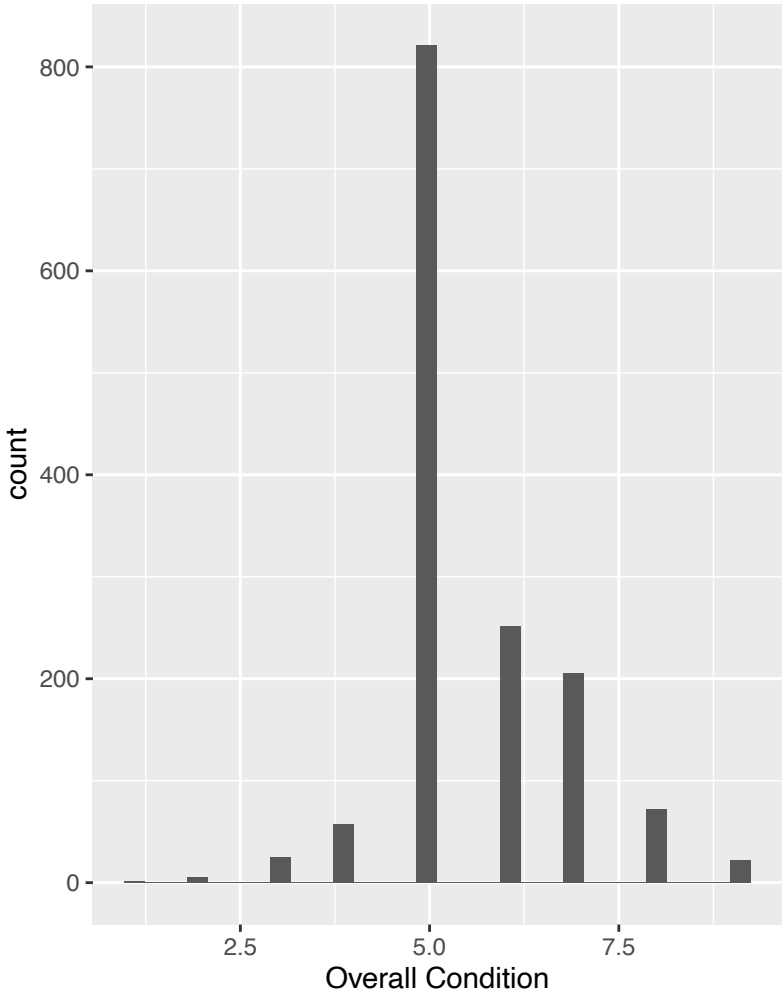
Style	Count
0.5Fin	154
1.5Unf	14
1Story	726
2.5Fin	8
2.5Unf	11
2Story	445
SFover	37
SLvl	65

3. Overall Quality



5.0 Average에 대부분 분포되어 있습니다.

4. Overall Condition



5.0 Average에 분포되어 있습니다.

1. Year Built

Min.	Q.1	Median	Mean	Q.3	Max.
1872	1954	1973	1971	2000	2010

Mode: 2006

2. YearRemodAdd: Remodeling Date

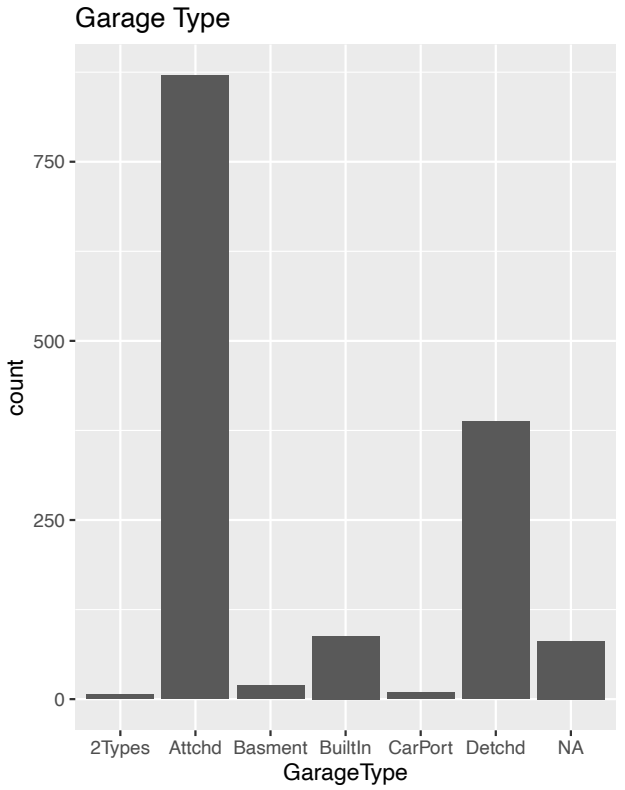
Min.	Q.1	Median	Mean	Q.3	Max.
1950	1967	1994	1985	2004	2010

Mode: 1950

3. Garage Type

`train$GarageType` count

* <chr>	<int>	
1 2Types	6	- more than one type of garage
2 Attchd	870	
3 Basment	19	
4 BuiltIn	88	
5 CarPort	9	
6 Detchd	387	
7 NA	81	- No garage



1. Number of Bedroom – 방 갯수 (지하 불포함)

Min.	Q.1	Median	Mean	Q.3	Max.
0.000	2.000	3.000	2.866	3.000	8.000

Mode: 3

2. Living Area SF – 거실 크기

Min.	Q.1	Median	Mean	Q.3	Max.
334	1130	1464	1515	1777	5642

Mode: 864

3. Total Rooms Above Ground – 방 갯수 (화장실 불포함)

Min.	Q.1	Median	Mean	Q.3	Max.
2.000	5.000	6.000	6.518	7.000	14.000

Mode: 6

- 지하를 포함하면 방의 갯수가 3개가 늘어나는 것으로 보아, 지하에는 평균적으로 3 bedrooms

1. Sale type

train$SaleType` count

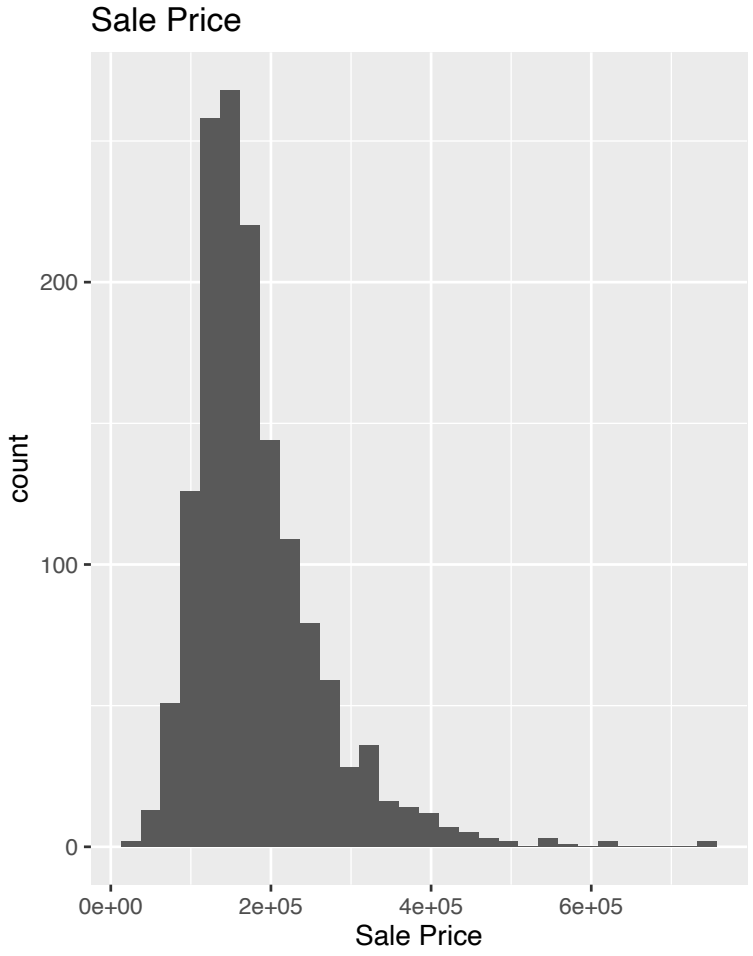
* <chr>	<int>
1 COD	43 – Court Officer Deed / Estate
2 Con	2 – 15% down payment regular terms
3 ConLD	9 – Contract Low Down
4 ConLI	5 – Contract Low Interest
5 ConLw	5 – Low Down & Low Interest
6 CWD	4 – Cash Warranty
7 New	122 - 새집
8 Oth	3
9 WD	1267 – Conventional Warranty

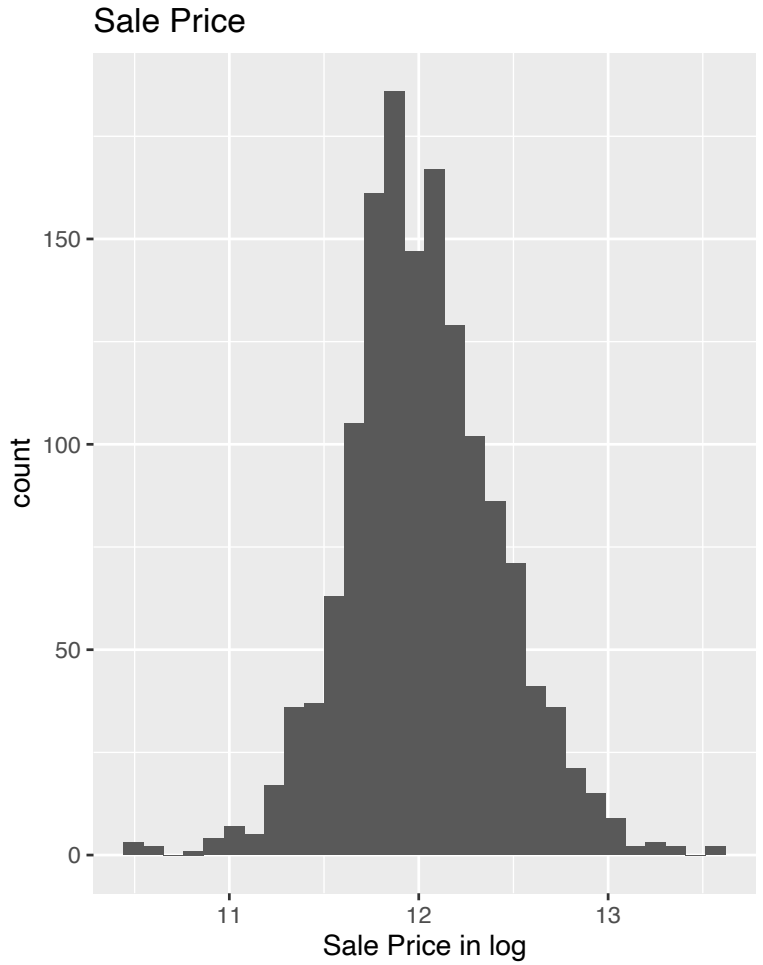
2. Sale Condition

`train$SaleCondition` count

* <chr>	<int>
1 Abnorml	101 – abnormal sale
2 AdjLand	4 – adjoining land purchase
3 Alloca	12 – condo (allocation)
4 Family	20 – 가족간의 매매
5 Normal	1198 – Normal
6 Partial	125 – 새집 지어지기 전

- Sale Price 분포도





Mean: $180,921
Median: $163,000
Mode: $140,000
Range: $34,900 ~ $755,000

주어져 있는 수치로 바로 그래프를 그리면 한쪽으로 치우쳐져 있어 정확한 수치를 얻을 수 없어서 전체 가격 수치 값에 로그를 걸어 데이터를 정규화 시켰습니다.

데이터 분석 방향

- Train 데이터는 집의 형태와 지어진 년도, 집의 구조, 형태에 따라 다르게 측정되어 있는 집의 매매 가격이 주어져 있는 데이터입니다. 그래서 집의 구조에 따라 가격이 어떻게 다른지, 집이 지어진 년도에 따라 가격이 어떻게 다른지에 대해 데이터 분석을 해나갈 예정입니다.

- Sales price 와 속성들과의 regression 으로 신뢰수준을 확인하고 정확도가 얼마나 되는지 분석
- 1. sales price ~ sale type
- 2. sales price ~ sale condition
- 3. sales price ~ number of total rooms
- 4. sales price ~ living room in sqft
- 5. sales price ~ year built (remodeling)
- (나중에 더 추가)