Name: Po Kit Man

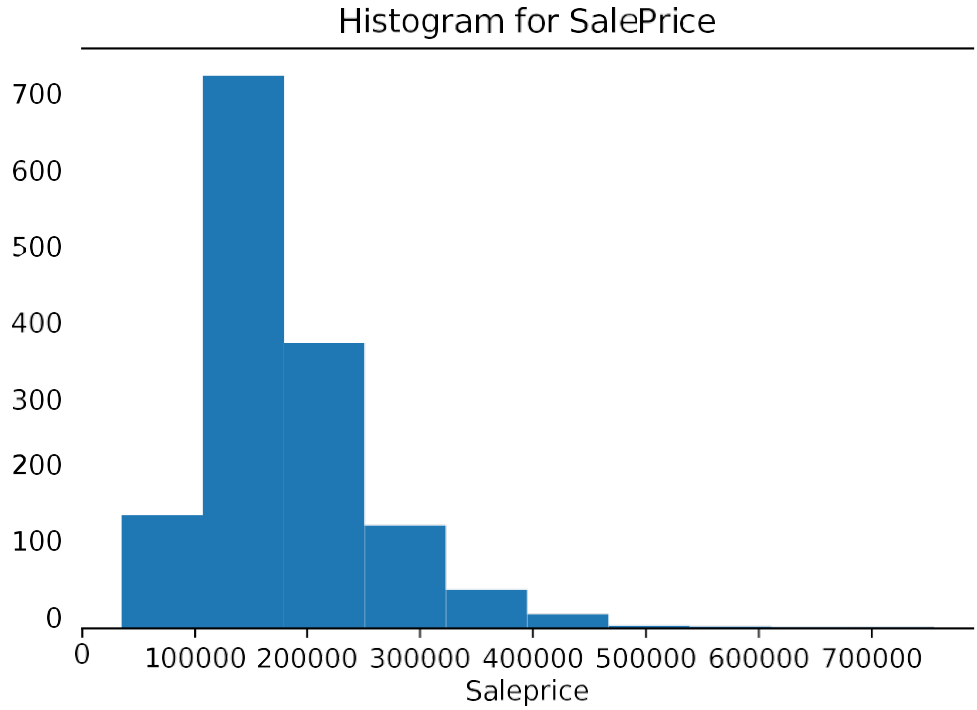
(Q1)

Nominal: [Id], [MSSubClass], [MSZoning], [Street], [Alley], [LotConfig], [Neighborhood], [Condition1], [Condition2], [BldgType], [HouseStyle], [RoofStyle], [RoofMatl], [Exterior1st], [Exterior2nd], [MasVnrType], [Foundation], [Heating], [CentralAir], [Electrical], [GarageType], [MiscFeature], [SaleType], [SaleCondition]

Ordinal: [LotShape], [LandContour], [Utilities], [LandSlope], [OverallQual], [OverallCond], [ExterQual], [ExterCond], [BsmtQual], [BsmtCond], [BsmtExposure], [BsmtFinType1], [BsmtFinType2], [HeatingQC], [KitchenQual], [Functional], [FireplaceQu], [GarageFinish], [GarageQual], [GarageCond], [PavedDrive], [PoolQC], [Fence]

Numeric: [LotFrontage], [LotArea], [YearBuilt], [YearRemodAdd], [MasVnrArea], [BsmtFinSF1], [BsmtFinSF2], [BsmtUnfSF], [TotalBsmtSF], [1stFlrSF], [2ndFlrSF], [LowQualFinSF], [GrLivArea], [BsmtFullBath], [BsmtHalfBath], [FullBath], [HalfBath], [BedroomAbvGr], [KitchenAbvGr], [TotRmsAbvGrd], [Fireplaces], [GarageYrBlt], [GarageCars], [GarageArea], [WoodDeckSF], [OpenPorchSF], [EnclosedPorch], [3SsnPorch], [ScreenPorch], [PoolArea], [MiscVal], [MoSold], [YrSold], [SalePrice]

(Q2)



(Q4) number of deleted records = 829

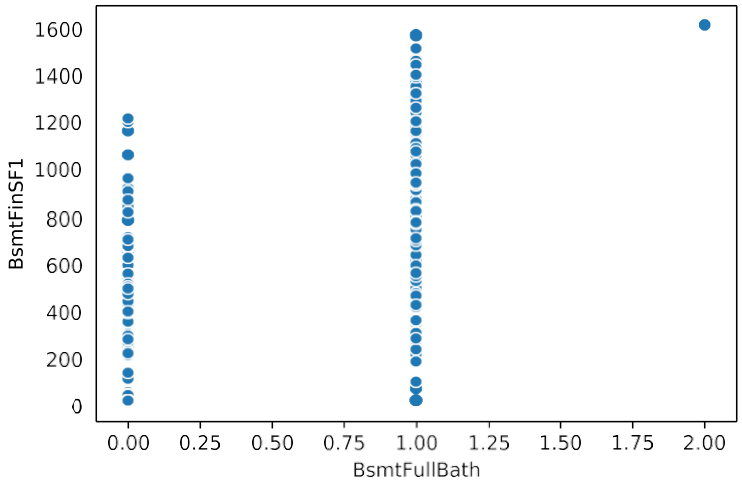
(Q5)
Removed attributes: [BsmtFinSF2], [LowQualFinSF], [BsmtHalfBath], [KitchenAbvGr], [EnclosedPorch], [3SsnPorch], [ScreenPorch], [PoolArea], [MiscVal]

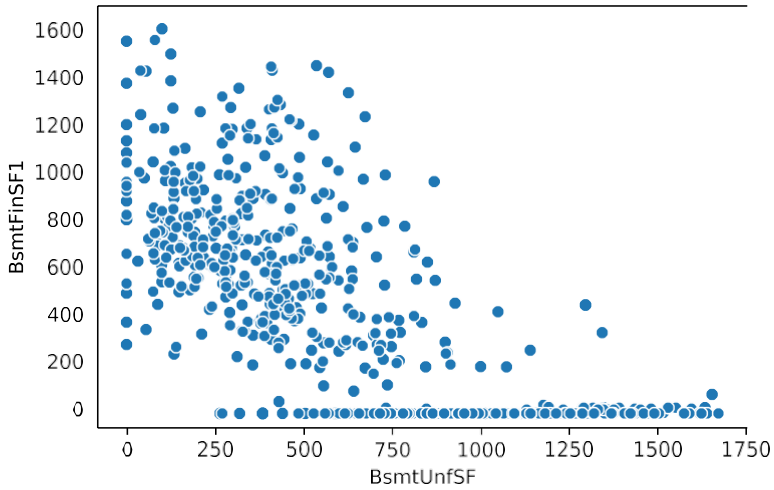
(Q6)

Top 5 numeric attributes that are most correlated with attribute ‘BsmtFinSF1’:

Attribute	Absolute value of correlation coefficient
BsmtFullBath	0.727655
BsmtUnfSF	0.712166
TotalBsmtSF	0.377109
1stFlrSF	0.351901
2ndFlrSF	0.231519

Numeric attributes ‘BsmtFullBath’ and ‘BsmtUnfSF’ will be removed because they have the two highest absolute value of correlation coefficient with ‘BsmtFinSF1’. This is to avoid the problem of multicollinearity. In addition, high correlation coefficients mean both of them can be predicted by the ‘BsmtFinSF1’ attribute. Thus, they are less important attributes comparing to the remaining three attributes.





(Q7)(a)
H0: Alley and GarageQual are independent
H1: Alley and GarageQual are not independent

Observed Frequency

Alley	Fa	Gd	NA	Po	TA	Total
Grvl	2	0	2	0	8	12
NA	15	2	26	1	562	606
Pave	1	0	0	0	12	13
Total	18	2	28	1	582	631

$$\begin{aligned} e_{ij} &= N \times P(A = a_i \wedge B = b_j) \\ &= N \times P(A = a_i) \times P(B = b_j) \\ &= \frac{1}{N}(count(A = a_i) \times count(B = b_j)) \end{aligned}$$

Expected Frequency

Alley	Fa	Gd	NA	Po	TA
Grvl	0.342314	0.038035	0.532488	0.019017	11.06815
NA	17.28685	1.920761	26.89065	0.96038	558.9414
Pave	0.37084	0.041204	0.576862	0.020602	11.99049

Chi-squared value for every cell

Alley	Fa	Gd	NA	Po	TA
Grvl	8.027499	0.038035	4.044393	0.019017	0.850505
NA	0.302523	0.003269	0.029499	0.001634	0.016737
Pave	1.067421	0.041204	0.576862	0.020602	7.54E-06

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

$\chi 2$ = 8.027499 + 0.302523 + 1.067421 + 0.038035 + 0.003269 + 0.041204 + 4.044393 + 0.029499 + 0.576862 + 0.019017 + 0.001634 + 0.020602 + 0.850505 + 0.016737 + 7.54E-06 = 15.03921

Chi-square value $\chi 2$ = 15.03921

Degree of freedom (df) = (5-1)*(3-1) = 8

Significance level = 0.001

From $\chi 2$ table, the critical value = 26.13

Since 15.03921 < 26.13, we will not reject the null hypothesis H0. 'Alley' is not dependent on 'GarageQual'. The 'Alley' attribute will not be removed.

(Q7)(b)
H0: BldgType and GarageQual are independent
H1: BldgType and GarageQual are not independent

Observed Frequency

BldgType	Fa	Gd	NA	Po	TA	Total
1Fam	18	2	27	1	502	550
2fmCon	0	0	0	0	5	5
Duplex	0	0	1	0	4	5
Twnhs	0	0	0	0	10	10
TwnhsE	0	0	0	0	61	61
Total	18	2	28	1	582	631

$$\begin{aligned} e_{ij} &= N \times P(A = a_i \wedge B = b_j) \\ &= N \times P(A = a_i) \times P(B = b_j) \\ &= \frac{1}{N}(count(A = a_i) \times count(B = b_j)) \end{aligned}$$

Expected Frequency

BldgType	Fa	Gd	NA	Po	TA
1Fam	15.68938	1.743265	24.40571	0.871632	507.29
2fmCon	0.142631	0.015848	0.22187	0.007924	4.611727
Duplex	0.142631	0.015848	0.22187	0.007924	4.611727
Twnhs	0.285261	0.031696	0.44374	0.015848	9.223455
TwnhsE	1.740095	0.193344	2.706815	0.096672	56.26307

Chi-squared value for every cell

BldgType	Fa	Gd	NA	Po	TA
1Fam	0.340291	0.03781	0.27577	0.018905	0.055164
2fmCon	0.142631	0.015848	0.22187	0.007924	0.03269
Duplex	0.142631	0.015848	2.729013	0.007924	0.081143
Twnhs	0.285261	0.031696	0.44374	0.015848	0.065379
TwnhsE	1.740095	0.193344	2.706815	0.096672	0.398813

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

$\chi 2$ = 0.340291 + 0.142631 + 0.142631 + 0.285261 + 1.740095 + 0.03781 + 0.015848 + 0.015848 + 0.031696 + 0.193344 + 0.27577 + 0.22187 + 2.729013 + 0.44374 + 2.706815 + 0.018905 + 0.007924 + 0.015848 + 0.096672 + 0.055164 + 0.03269 + 0.081143 + 0.065379 + 0.398813 = 10.10312

Chi-square value $\chi 2$ = 10.10312

Degree of freedom (df) = (5-1)*(5-1) = 16

Significance level = 0.001

From $\chi 2$ table, the critical value = 39.25

Since 10.10312 < 39.25, we will not reject the null hypothesis H0. ‘BldgType’ is not dependent on ‘GarageQual’. The ‘BldgType’ attribute will not be removed.

(Q7)(c)
H0: GarageCond and GarageQual are independent
H1: GarageCond and GarageQual are not independent

Observed Frequency

GarageCond	Fa	Gd	NA	Po	TA	Total
Fa	7	0	0	0	1	8
Gd	0	0	0	0	2	2
NA	0	0	28	0	0	28
Po	1	0	0	1	0	2
TA	10	2	0	0	579	591
Total	18	2	28	1	582	631

$$\begin{aligned} e_{ij} &= N \times P(A = a_i \wedge B = b_j) \\ &= N \times P(A = a_i) \times P(B = b_j) \\ &= \frac{1}{N}(count(A = a_i) \times count(B = b_j)) \end{aligned}$$

Expected Frequency

GarageCond	Fa	Gd	NA	Po	TA
Fa	0.228209	0.025357	0.354992	0.012678	7.378764
Gd	0.057052	0.006339	0.088748	0.00317	1.844691
NA	0.798732	0.088748	1.242472	0.044374	25.82567
Po	0.057052	0.006339	0.088748	0.00317	1.844691
TA	16.85895	1.873217	26.22504	0.936609	545.1062

Chi-squared value for every cell

GarageCond	Fa	Gd	NA	Po	TA
Fa	200.9435	0.025357	0.354992	0.012678	5.514288
Gd	0.057052	0.006339	0.088748	0.00317	0.013076
NA	0.798732	0.088748	576.2425	0.044374	25.82567
Po	15.58483	0.006339	0.088748	313.5032	1.844691
TA	2.79052	0.008581	26.22504	0.936609	2.107463

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

$\chi 2$ = 200.9435 + 0.057052 + 0.798732 + 15.58483 + 2.79052 + 0.025357 + 0.006339 + 0.088748 + 0.006339 + 0.008581 + 0.354992 + 0.088748 + 576.2425 + 0.088748 + 26.22504 + 0.012678 + 0.00317 + 0.044374 + 313.5032 + 0.936609 + 5.514288 + 0.013076 + 25.82567 + 1.844691 + 2.107463 = 1173.115

Chi-square value $\chi 2$ = 1173.115

Degree of freedom (df) = (5-1)*(5-1) = 16

Significance level = 0.001

From $\chi 2$ table, the critical value = 39.25

Since 1173.115 > 39.25, we will reject the null hypothesis H0. ‘GarageCond’ and ‘GarageQual’ are not independent. The ‘GarageCond’ attribute will be removed.

(Q8)

Mean values for each numeric attribute

‘LotFrontage’: 67.98046875

‘MasVnrArea’: 76.1150159744409

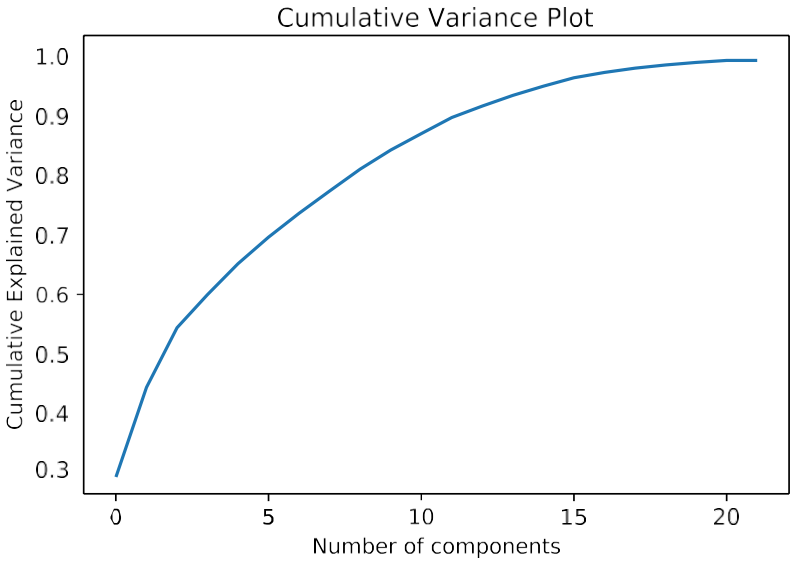
‘GarageYrBlt’: 1985.5373134328358

(Q9)

For attribute ‘MaxVnrType’, the filled-in value = ‘None’

For attribute ‘Electrical’, the filled-in value = 'SBrkr'

(Q11)



The number of smallest set of PCA features is 12.

	min	25%	50%	75%	max
PCA1	-6.72057	-2.02548	0.361374	1.745746	7.167636
PCA2	-4.00655	-1.48018	-0.15398	1.697272	4.817172
PCA3	-4.55419	-1.06163	0.002411	1.051319	4.578652
PCA4	-2.99849	-0.71381	0.019017	0.804957	3.054826
PCA5	-3.08282	-0.75485	-0.06363	0.707279	3.419611
PCA6	-2.77155	-0.63511	0.000198	0.617376	3.058426
PCA7	-3.105	-0.58121	-0.00683	0.70247	3.218709
PCA8	-2.40856	-0.62407	-0.05137	0.603462	2.950301
PCA9	-2.93759	-0.58134	0.015165	0.590822	3.020794
PCA10	-2.59273	-0.55822	0.001992	0.581314	3.15809
PCA11	-2.46129	-0.48398	0.001052	0.473934	2.778489
PCA12	-2.63563	-0.47748	-0.00564	0.476523	2.980374