Exploratory Analysis의 결과 및 분석 계획

- 2019.11.05 -

응용정보통계학과

김도혜

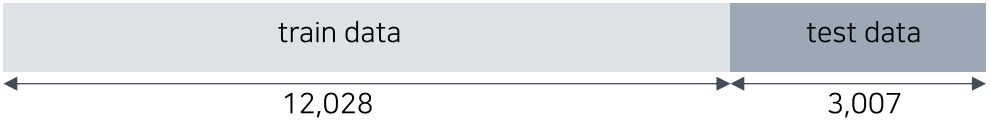
목차

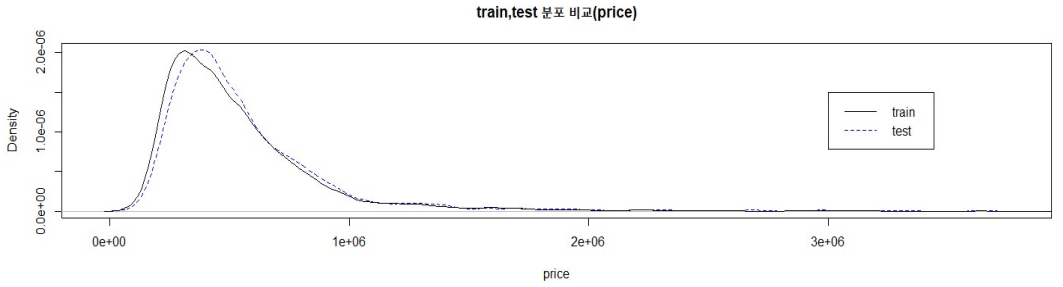
1. 분석 개요

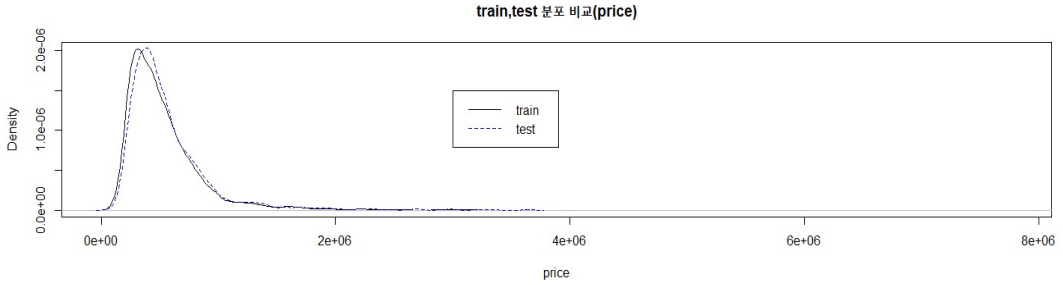
- 분석 목적 : 주택 가격에 영향을 미치는 요인을 찾고, 가격을 예측하기 위함

- 데이터의 구성 : 2014년 5월 ~ 2015년 5월까지 15,035개가 수집됨.
  id 변수를 제외하고 20개의 변수로 구성되어있음.
  목표 변수는 'price(가격)'이며, 나머지 19개의 변수는 이를 예측하기 위한 입력 변수들임.

- 예측 평가를 위해 train data : test data = 8 : 2로 데이터를 나눔





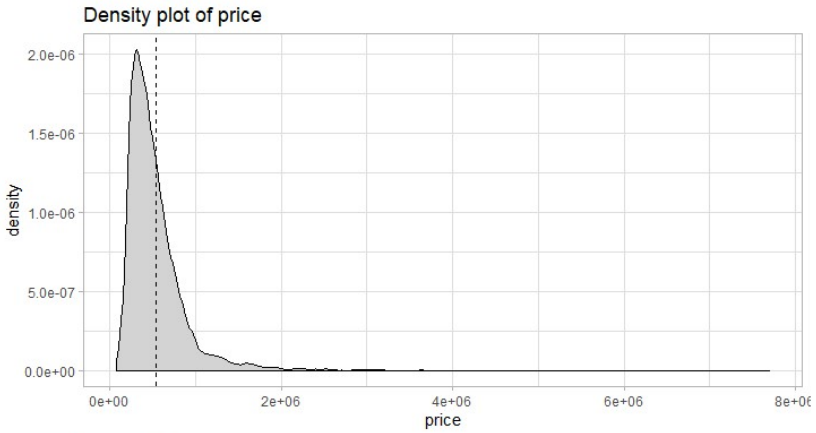


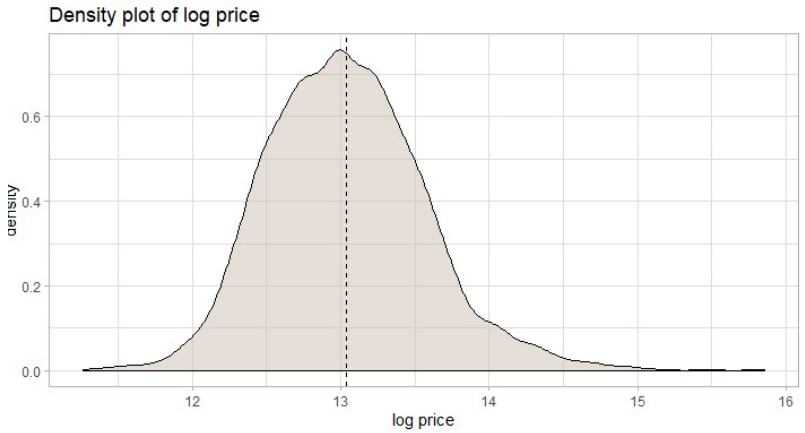
2. 목표 변수(price) 파악

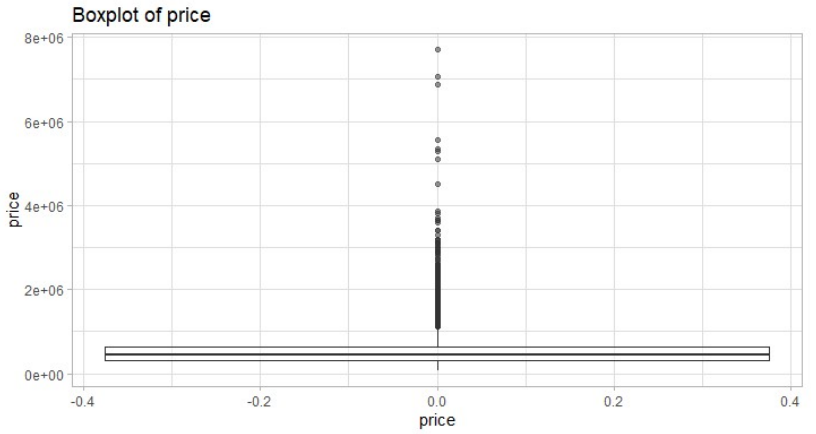
min.	median	max.	mean	sd
78000	448000	7700000	534776.4	371843.1

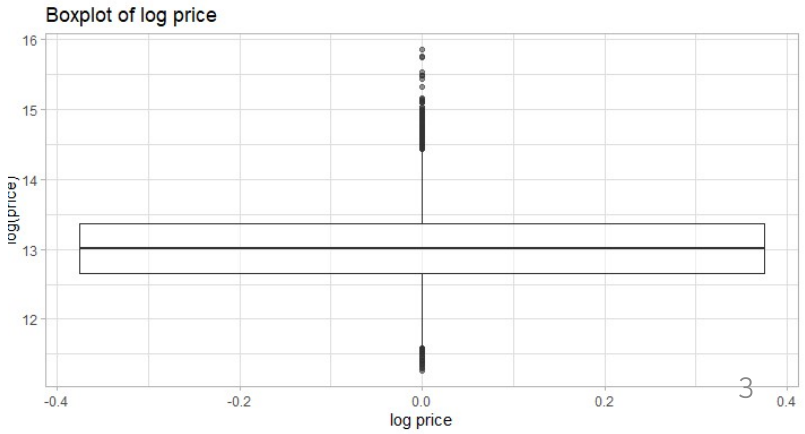
log 변환

min.	median	max.	mean	sd
11.264	13.012	15.857	13.035	0.531









3. 가격이 가장 높은/낮은 주택의 특징

- price 변수 기준, 상위 10개 주택

id	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_basement	yr_built	yr_renovated	zipcode	lat	long	sqft_living15	sqft_lot15
5108	7700000	6	8	12050	27600	2.5	0	3	4	13	8570	3480	1910	1987	98102	47.63	-122.32	3940	8800
2775	7062500	5	4.5	10040	37325	2	1	2	3	11	7680	2360	1940	2001	98004	47.65	-122.21	3930	25449
6469	6885000	6	7.75	9890	31374	2	0	4	3	13	8860	1030	2001	0	98039	47.631	-122.24	4540	42730
3134	5570000	5	5.75	9200	35069	2	0	0	3	13	6200	3000	2001	0	98039	47.629	-122.23	3560	24345
1045	5350000	5	5	8000	23985	2	0	4	3	12	6720	1280	2009	0	98004	47.623	-122.22	4600	21750
947	5300000	6	6	7390	24829	2	1	4	4	12	5000	2390	1991	0	98040	47.563	-122.21	4320	24619
842	5110800	5	5.25	8010	45517	2	1	4	3	12	5990	2020	1999	0	98033	47.677	-122.21	3430	26788
1882	4500000	5	5.5	6640	40014	2	1	4	3	12	6350	290	2004	0	98155	47.749	-122.28	3030	23408
1499	3850000	4	4.25	5770	21300	2	1	4	4	11	5770	0	1980	0	98040	47.585	-122.22	4620	22748
4957	3800000	5	5.5	7050	42840	1	0	2	4	13	4320	2730	1978	0	98004	47.623	-122.22	5070	20570
mean	5512830	5.2	5.75	8404	32985	1.95	0.5	3.1	3.4	12.2	6546	1858	1981	398.8	98046	47.636	-122.24	4104	24121

- 방의 수(bedrooms), 화장실 수(bathrooms)가 많음
- 주택의 면적(sqft_living)이 넓음(부지 면적(sqft_lot)은 대체로 넓으나, 비례하진 않는 것 같음)
- 집 앞에 강이 흐르는 집(waterfront=1)이 6곳이나 있었음 (전체 train data12,028개 중 90개만 '1'의 값을 가짐)
- 등급(grade)이 매우 높다. 상태(condition)도 대체로 좋은 편
- view는 다 높은 것 아님. 층수(floors)도 큰 영향은 아닐 듯함.
- 위도(lat)가 비교적 큰 값을 가짐 (시애틀의 집값은 북부가 더 비싸다고 함)
- 우편번호(zipcode) 98039, 98004, 98040가 유독 많이 보임

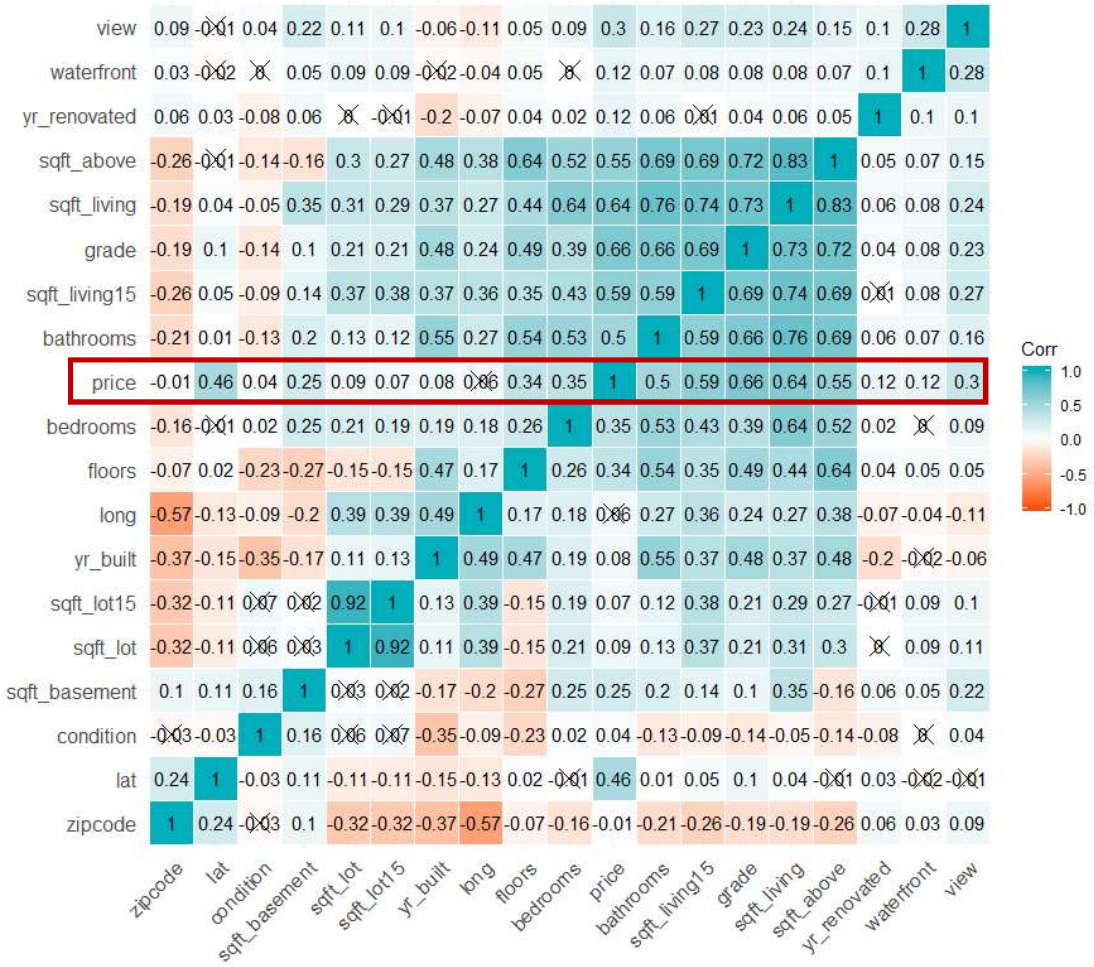
3. 가격이 가장 높은/낮은 주택의 특징

- price 변수 기준, 하위 10개 주택

id	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_basement	yr_built	yr_renovated	zipcode	lat	long	sqft_living15	sqft_lot15
10678	78000	2	1	780	16344	1	0	0	1	5	780	0	1942	0	98168	47.474	-122.28	1700	10387
339	80000	1	0.75	430	5050	1	0	0	2	4	430	0	1912	0	98014	47.65	-121.91	1200	7500
11293	81000	2	1	730	9975	1	0	0	1	5	730	0	1943	0	98168	47.481	-122.32	860	9000
2674	84000	2	1	700	20130	1	0	0	3	6	700	0	1949	0	98168	47.475	-122.27	1490	18630
7166	85000	2	1	830	9000	1	0	0	3	6	830	0	1939	0	98032	47.381	-122.24	1160	7680
11642	85000	2	1	910	9753	1	0	0	3	5	910	0	1947	0	98032	47.39	-122.24	1160	7405
4146	89000	3	1	900	4750	1	0	0	4	6	900	0	1969	0	98023	47.303	-122.36	900	3404
2219	89950	1	1	570	4080	1	0	0	3	5	570	0	1942	0	98146	47.51	-122.33	890	5100
5623	90000	1	1	780	4000	1	0	0	3	5	780	0	1905	0	98108	47.542	-122.32	1150	4000
8756	90000	2	1	790	2640	1	0	0	3	7	790	0	1973	0	98034	47.735	-122.18	1310	2064
mean	85195	1.8	0.975	742	8572.2	1	0	0	2.6	5.4	742	0	1942	0	98089	47.494	-122.25	1182	7517

- 방의 수(bedrooms), 화장실 수(bathrooms)가 적음
- 주택의 면적(sqft_living)이 좁음 (여기도 부지 면적(sqft_lot)은 비례하진 않는 것 같음)
- 집 앞에 강이 흐르는 집(waterfront=1)이 없음
- 등급(grade)이 대체로 낮거나 중간이고, 상태(condition)는 중간이거나 나쁜 정도임.
- view가 모두 0점이고, 층수(floors)가 모두 1층임.
- 위도(lat)가 대체로 낮은 값을 가짐
- 우편번호(zipcode) 98168, 98032가 많이 보임

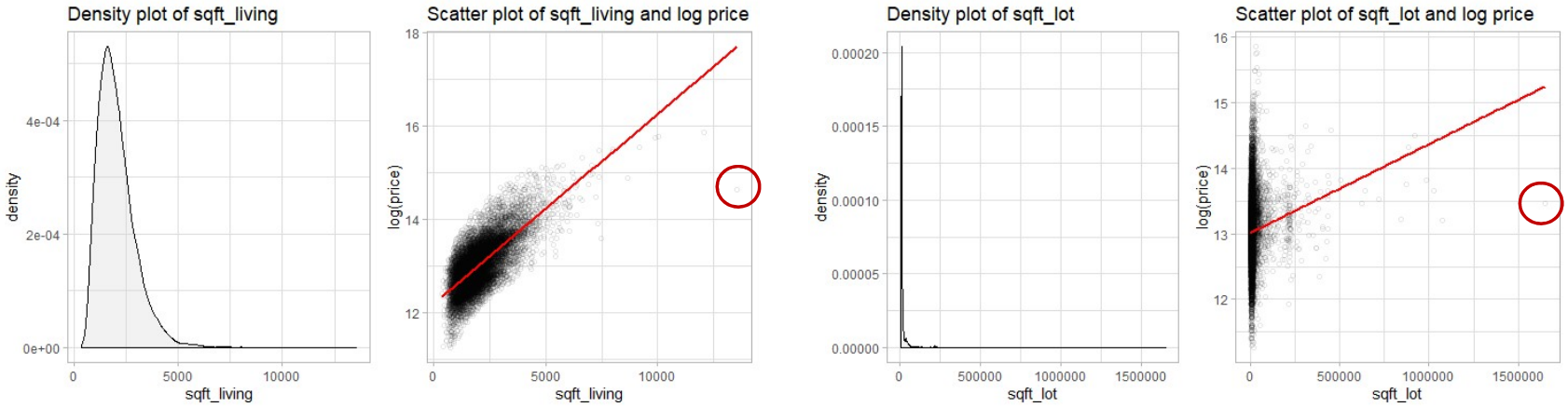
4. 변수들 간의 상관관계 및 변수 파악



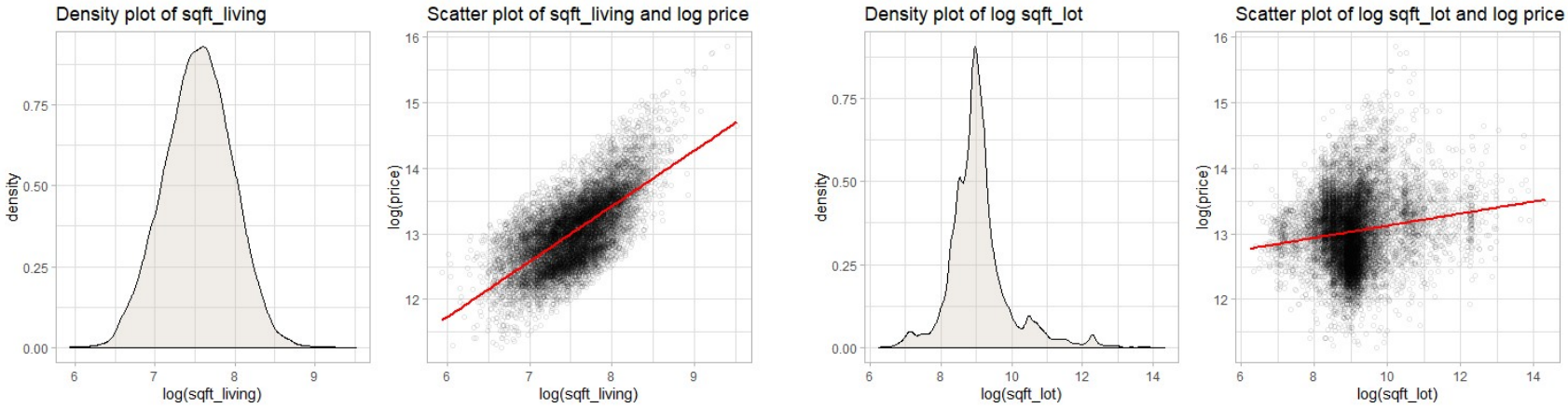
목적변수(price)와의 상관계수

1. grade : 0.66
2. sqft_living : 0.64
3. sqft_living15 : 0.59
4. sqft_above : 0.55
5. bathrooms : 0.5
6. lat : 0.46
7. bedrooms : 0.35
8. floors : 0.34
9. view : 0.30
10. sqft_basement : 0.25
11. yr_new : 0.19
12. yr_renovated : 0.12
13. waterfront : 0.12

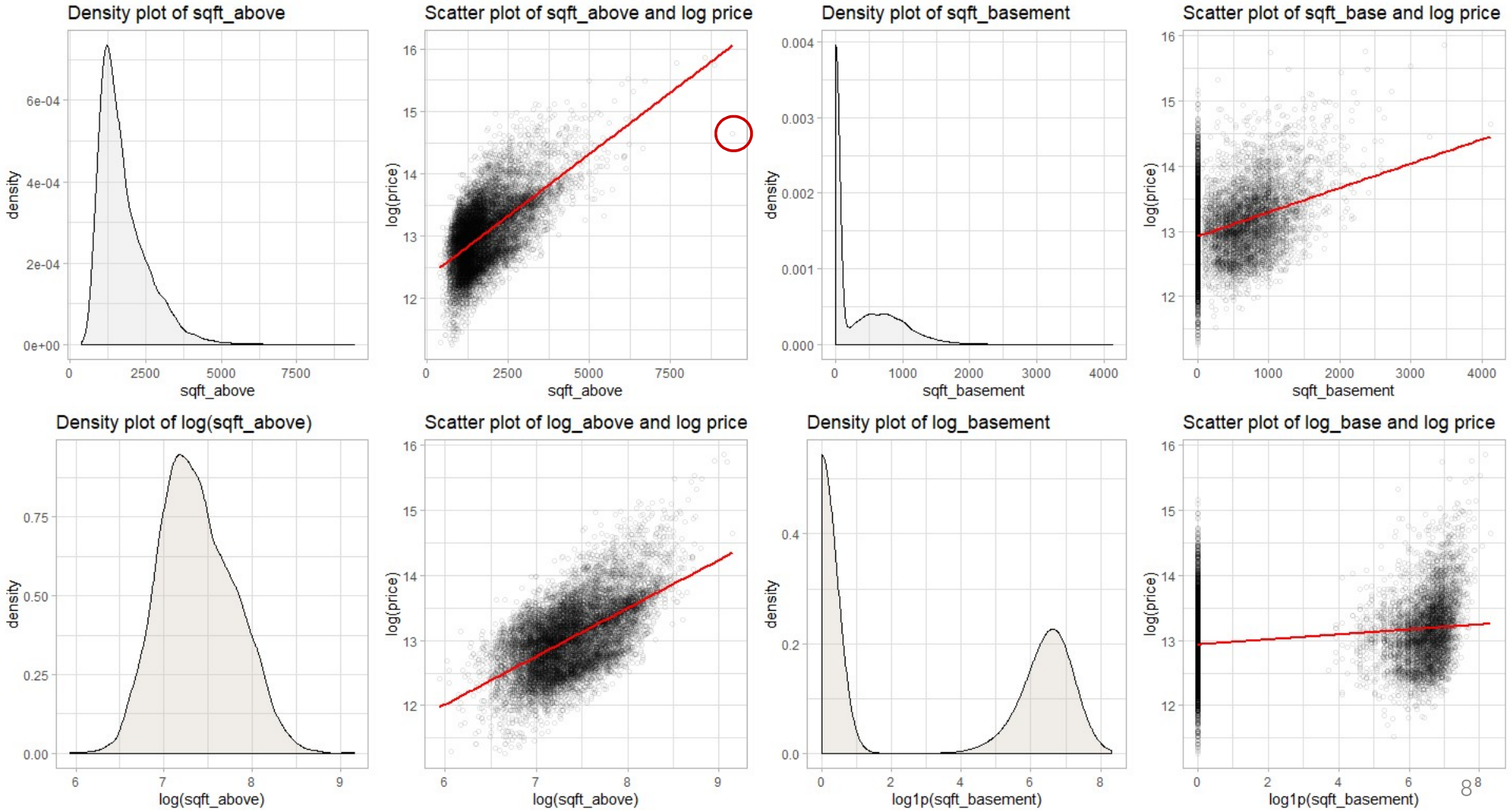
sqft_living & sqft_lot * Price와의 상관계수 : 0.64 & 0.09



log 변환

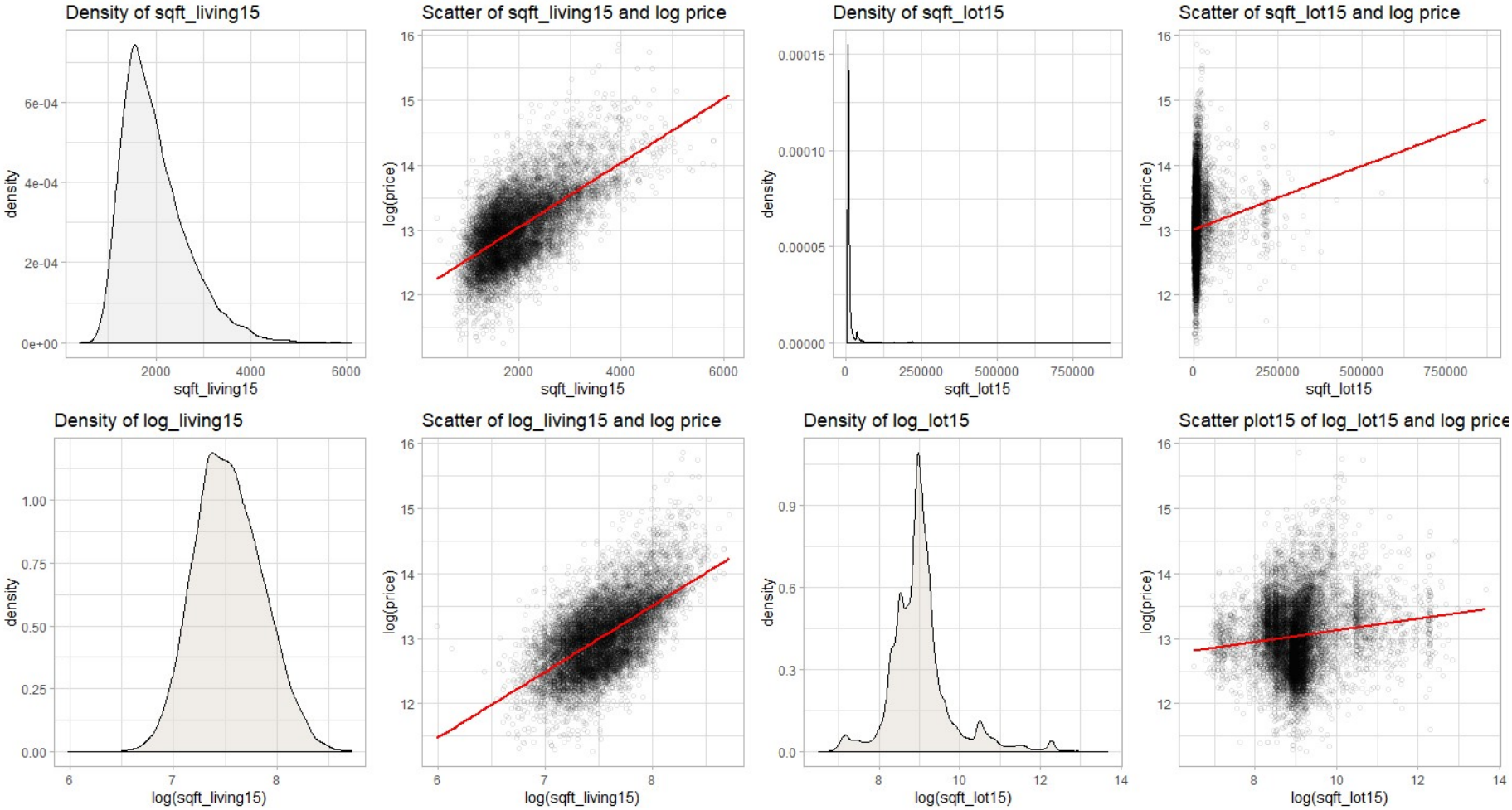


sqft_above & sqft_basement * Price와의 상관계수 : 0.55 & 0.25



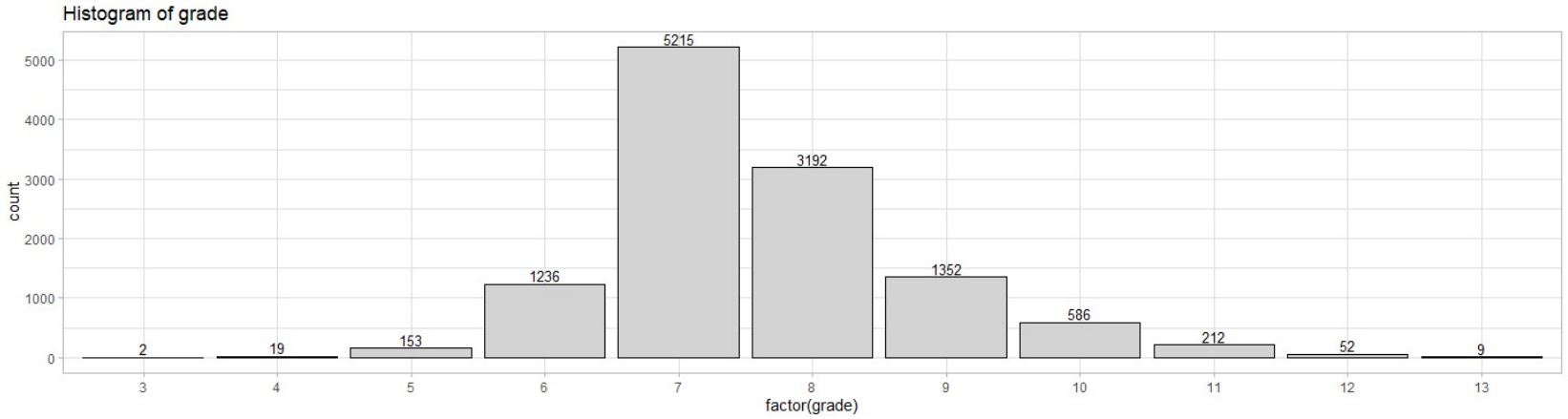
sqft_living15 & sqft_lot15

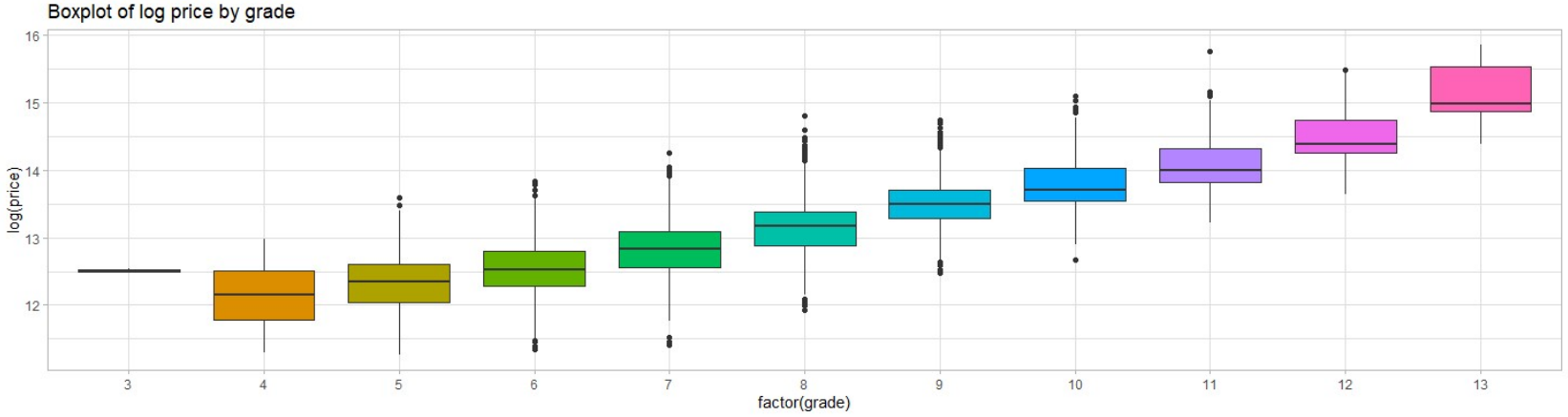
* Price와의 상관계수 : 0.59 & 0.07



grade * Price와의 상관계수 : 0.66

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
3	7	7	7.599	8	13





grade

BUILDING GRADE

Represents the construction quality of improvements. Grades run from grade 1 to 13. Generally defined as:

1-3 Falls short of minimum building standards. Normally cabin or inferior structure.

4 Generally older, low quality construction. Does not meet code.

5 Low construction costs and workmanship. Small, simple design.

6 Lowest grade currently meeting building code. Low quality materials and simple designs.

7 Average grade of construction and design. Commonly seen in plats and older sub-divisions.

8 Just above average in construction and design. Usually better materials in both the exterior and interior finish work.

9 Better architectural design with extra interior and exterior design and quality.

10 Homes of this quality generally have high quality features. Finish work is better and more design quality is seen in the floor plans. Generally have a larger square footage.

11 Custom design and higher quality finish work with added amenities of solid woods, bathroom fixtures and more luxurious options.

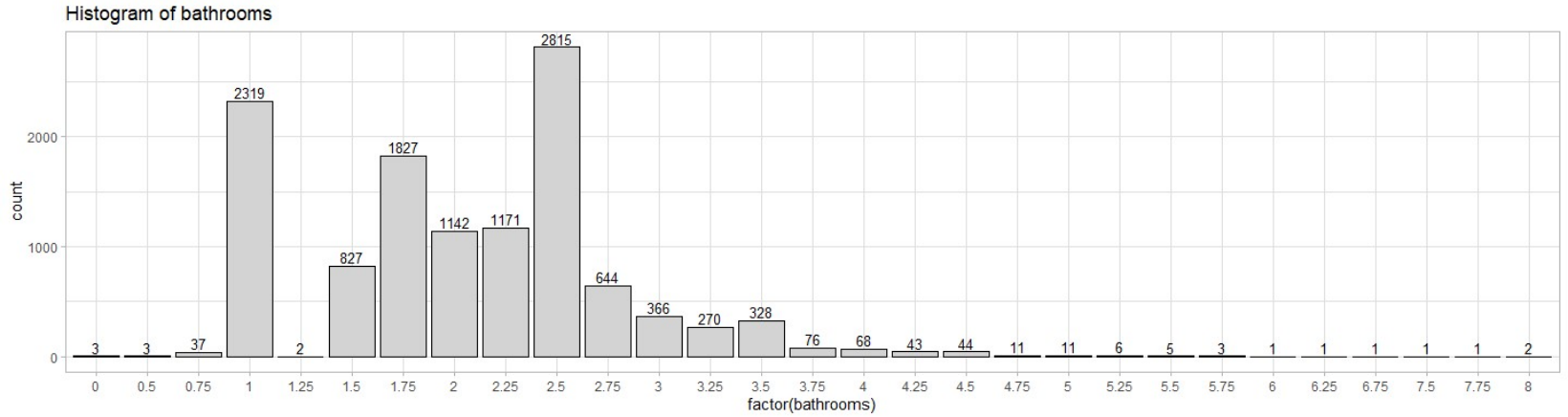
12 Custom design and excellent builders. All materials are of the highest quality and all conveniences are present.

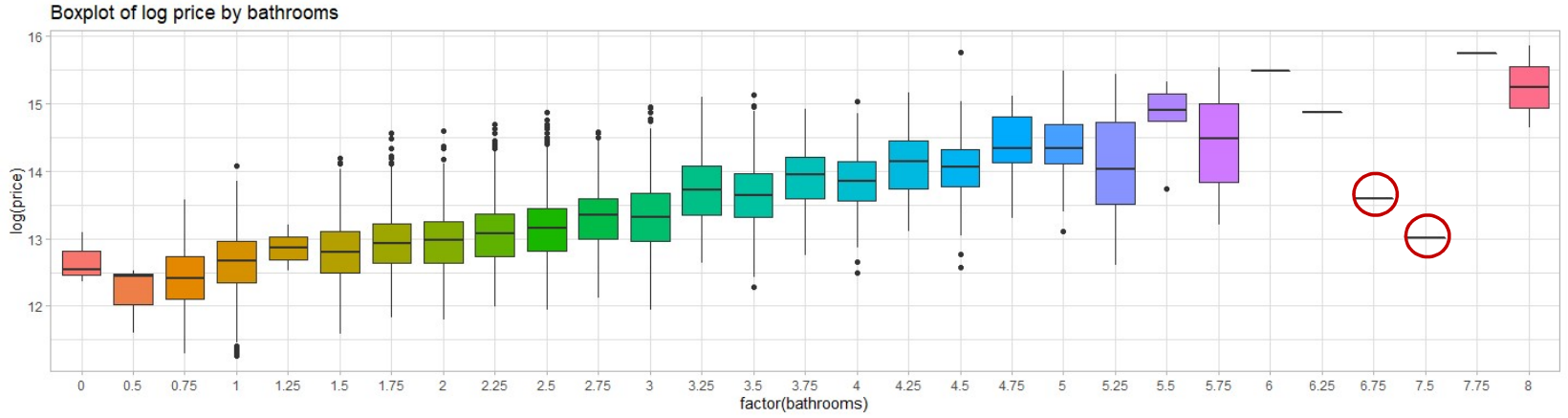
13 Generally custom designed and built. Mansion level. Large amount of highest quality cabinet work, wood trim, marble, entry ways etc.

https://info.kingcounty.gov/assessor/esales/Glossary.aspx?type=r

bathrooms * Price와의 상관계수 : 0.5

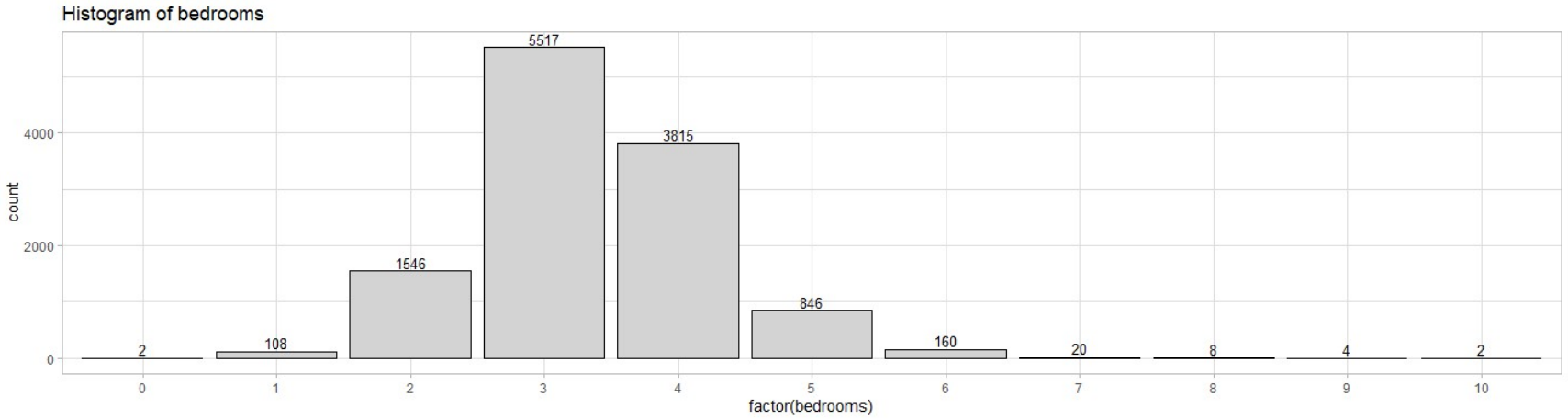
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	1.5	2.0	2.063	2.5	8.0

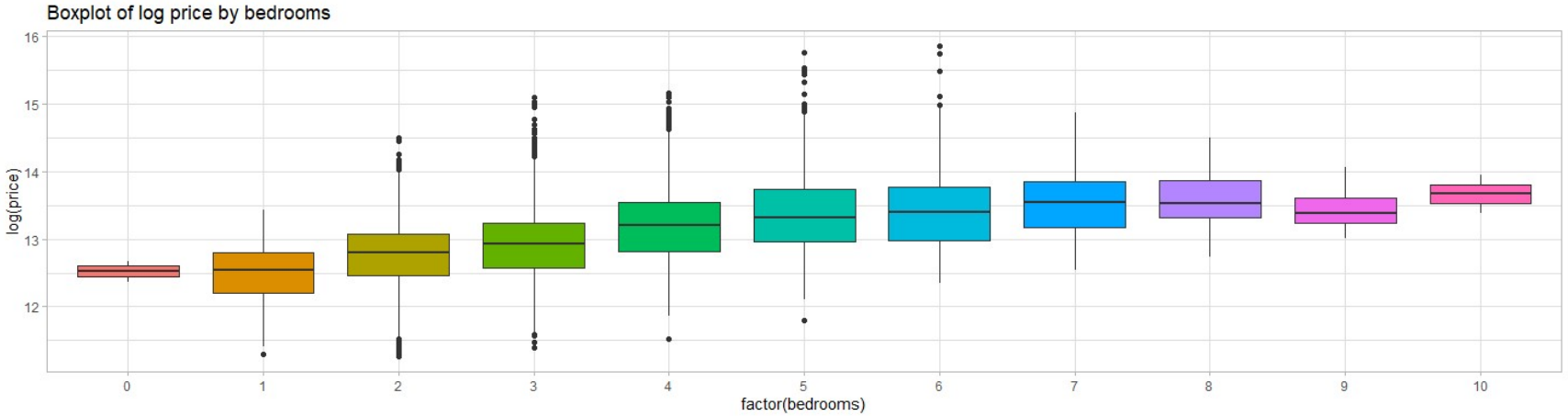




bedrooms * Price와의 상관계수 : 0.35

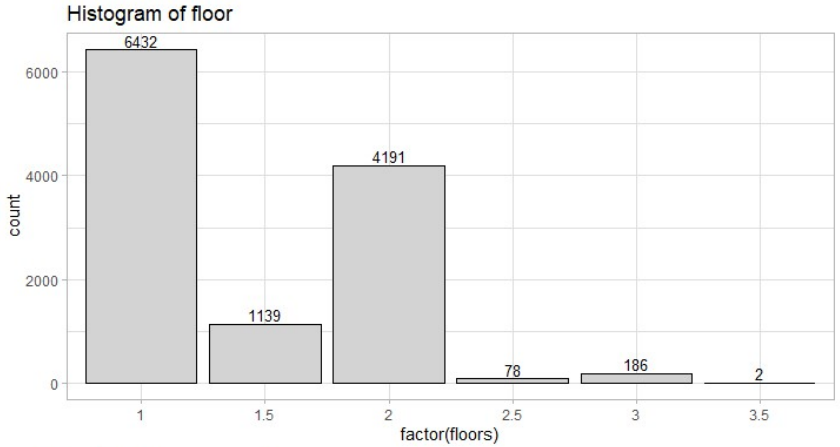
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	3	3	3.364	4	10

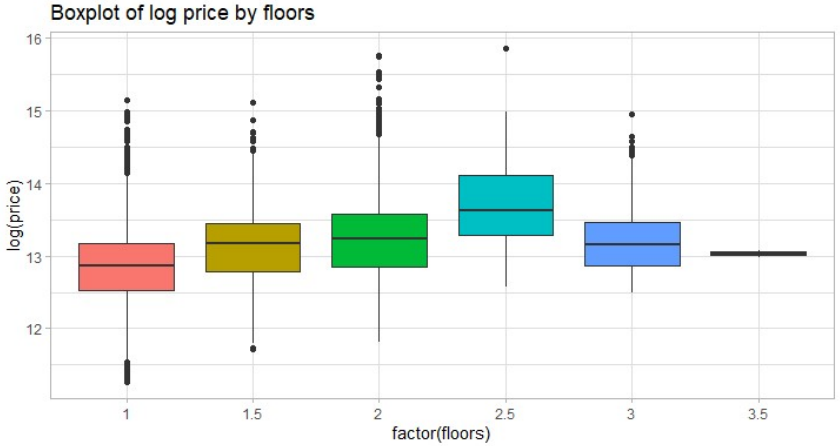




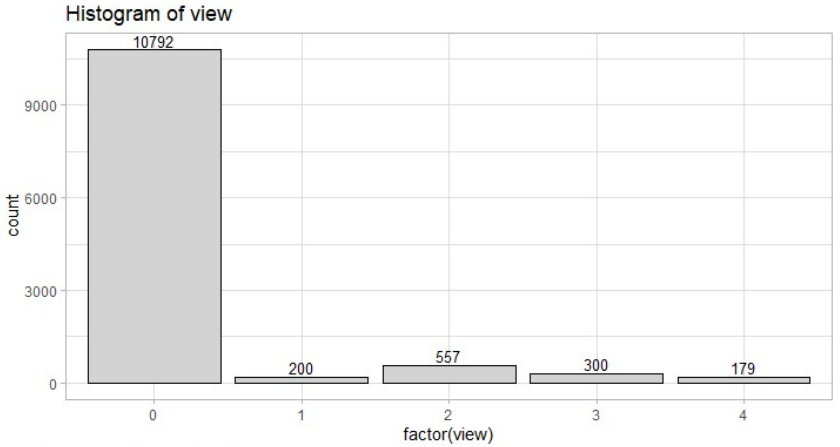
floor & view * Price와의 상관계수 : 0.34 & 0.30

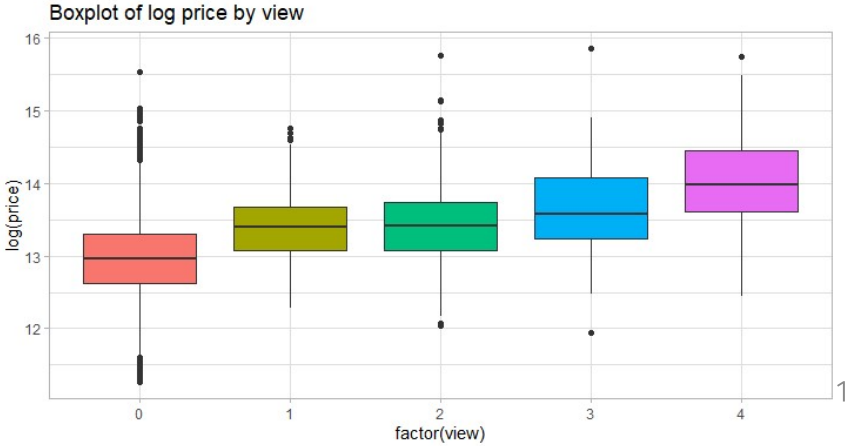
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	1	1	1.437	2	3.5





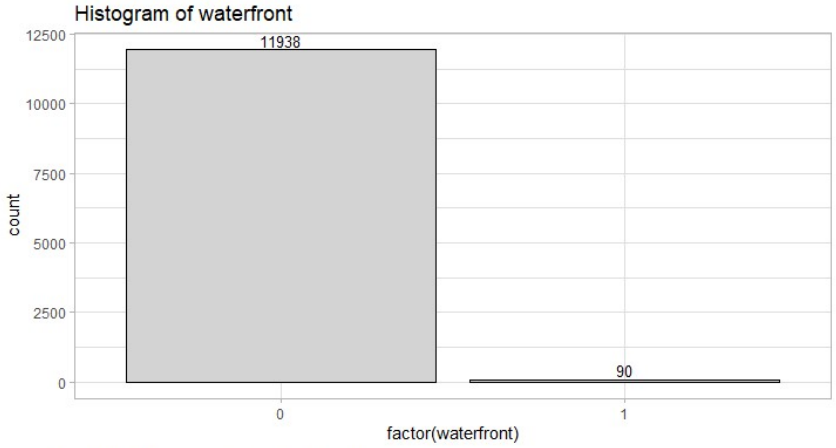
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	0	0	0.244	0	4

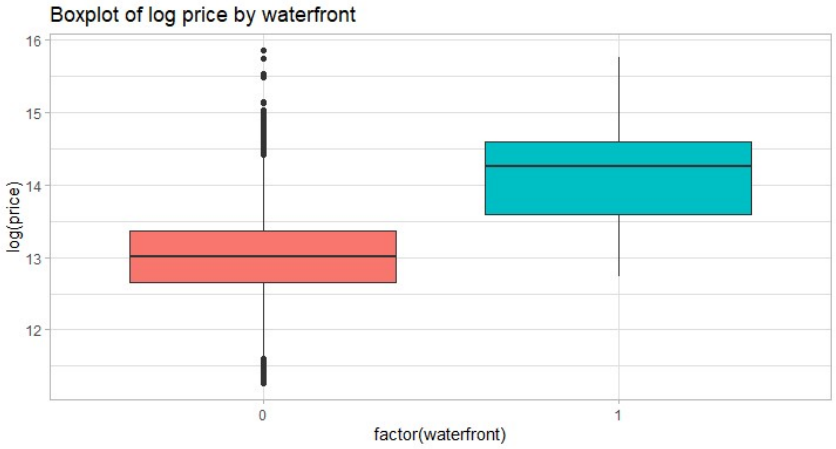




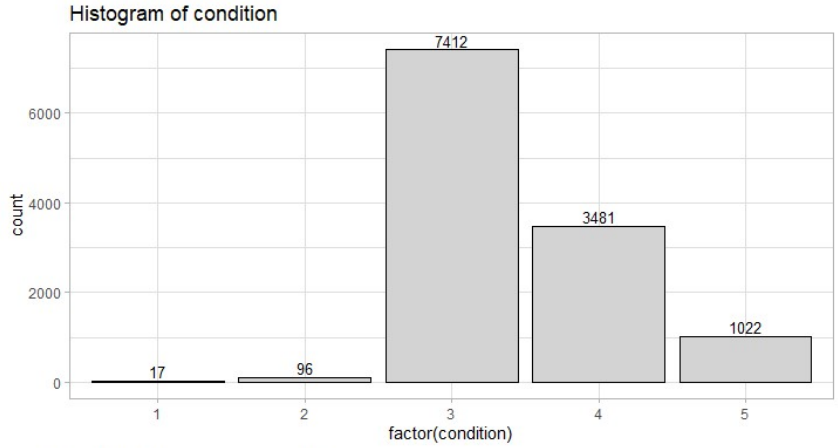
waterfront & condition * Price와의 상관계수 : 0.12 & 0.04

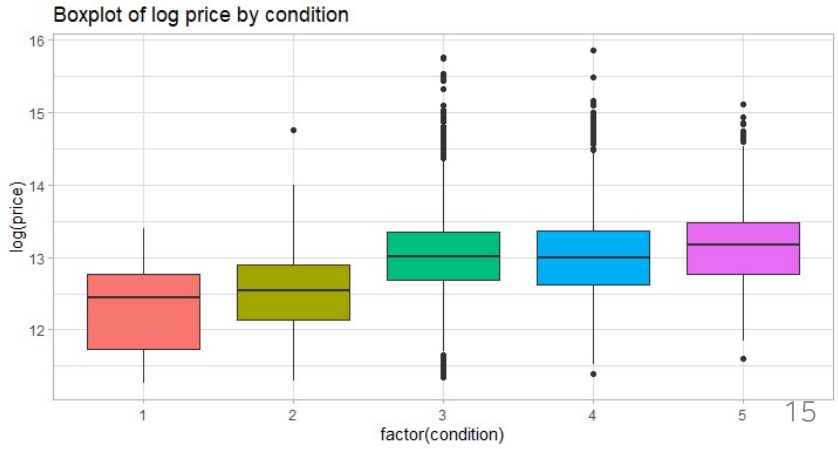
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	0	0	0.007	0	1





Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	3	3	3.449	4	5





condition

BUILDING CONDITION

Relative to age and grade. Coded 1-5.

1 = Poor- Worn out. Repair and overhaul needed on painted surfaces, roofing, plumbing, heating and numerous functional inadequacies. Excessive deferred maintenance and abuse, limited value-in-use, approaching abandonment or major reconstruction; reuse or change in occupancy is imminent. Effective age is near the end of the scale regardless of the actual chronological age.

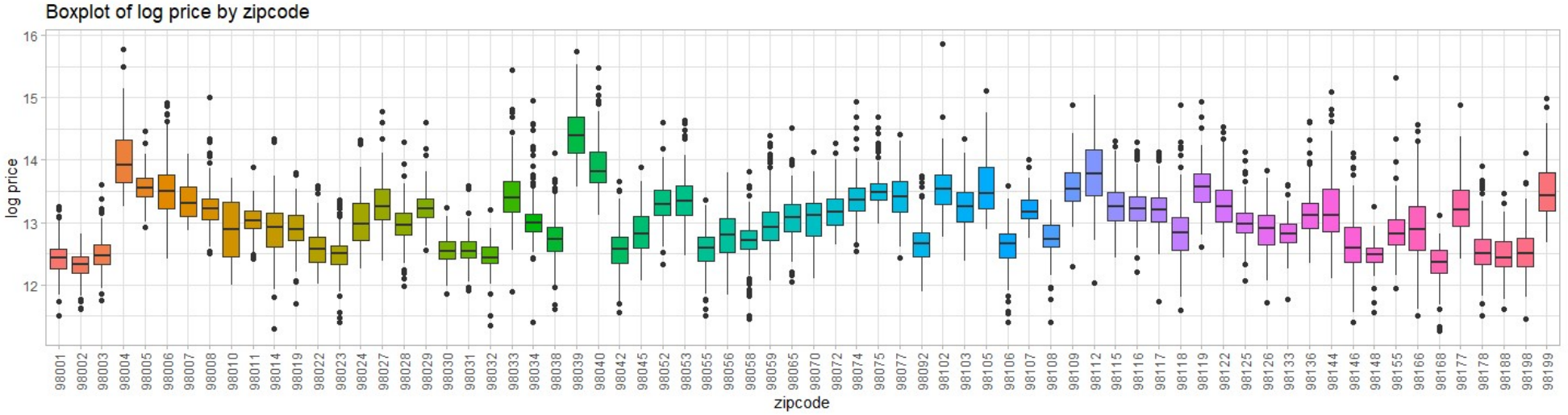
2 = Fair- Badly worn. Much repair needed. Many items need refinishing or overhauling, deferred maintenance obvious, inadequate building utility and systems all shortening the life expectancy and increasing the effective age.

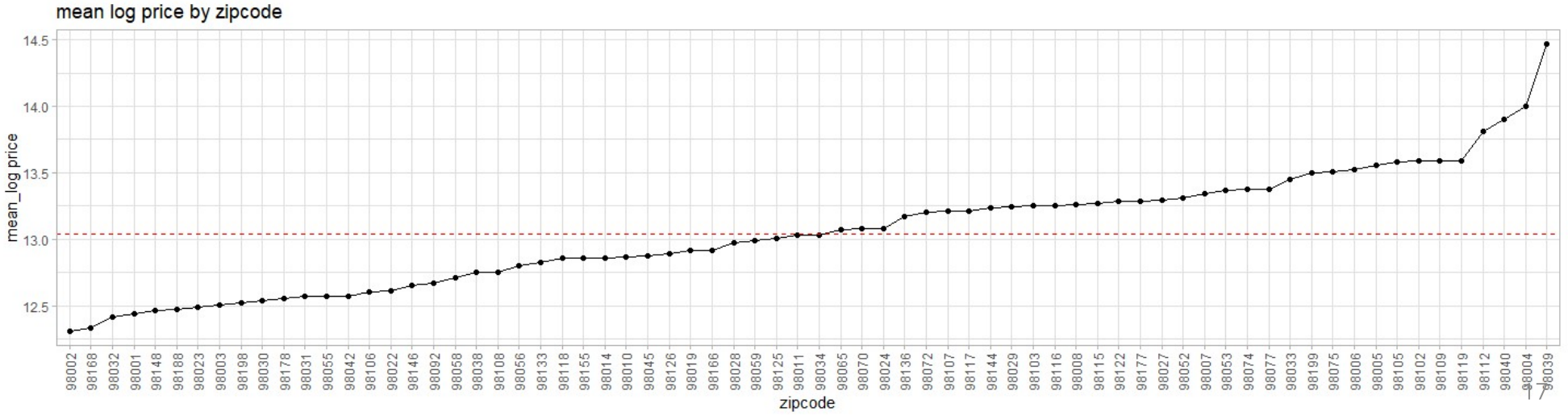
3 = Average- Some evidence of deferred maintenance and normal obsolescence with age in that a few minor repairs are needed, along with some refinishing. All major components still functional and contributing toward an extended life expectancy. Effective age and utility is standard for like properties of its class and usage.

4 = Good- No obvious maintenance required but neither is everything new. Appearance and utility are above the standard and the overall effective age will be lower than the typical property.

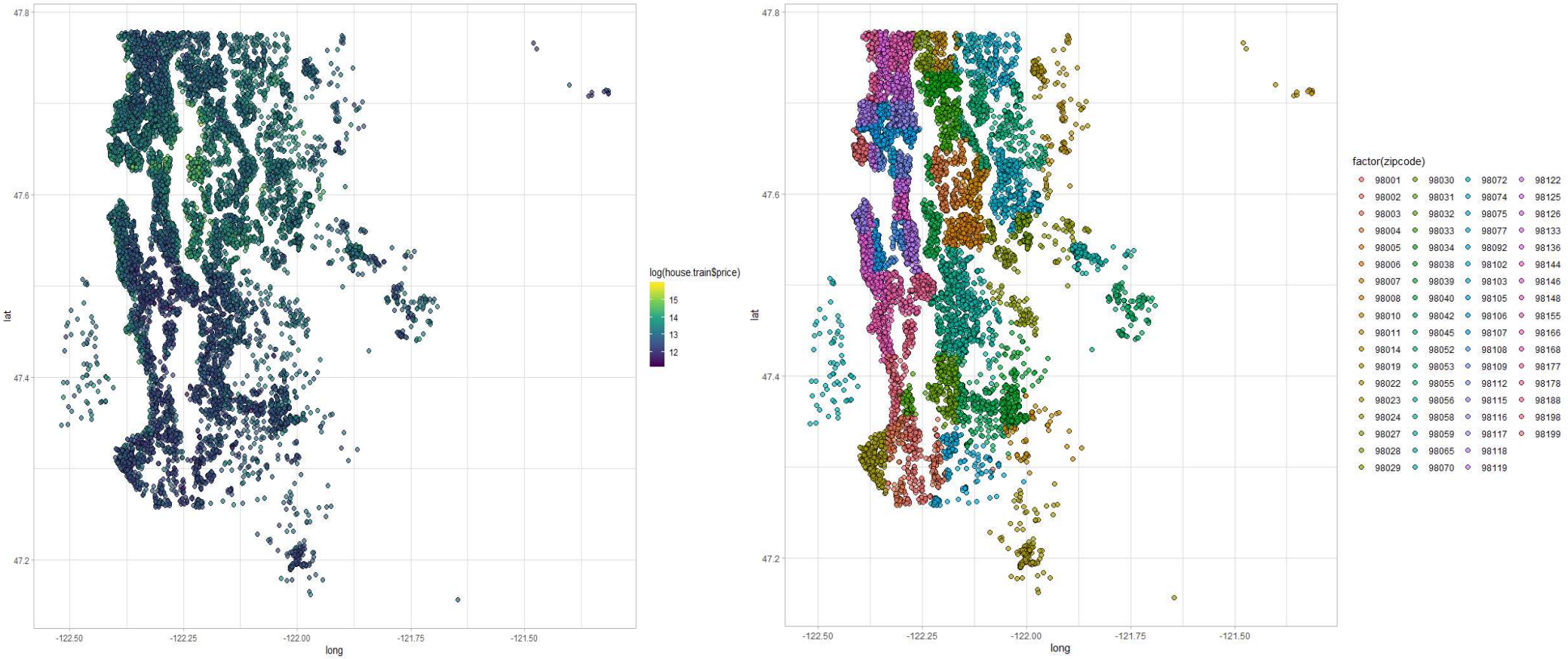
5= Very Good- All items well maintained, many having been overhauled and repaired as they have shown signs of wear, increasing the life expectancy and lowering the effective age with little deterioration or obsolescence evident with a high degree of utility.

zipcode


Boxplot of log price by zipcode
log price
16
15
14
13
12
98001 98002 98003 98004 98005 98006 98007 98008 98010 98011 98014 98019 98022 98023 98024 98027 98028 98029 98030 98031 98032 98033 98034 98038 98039 98040 98042 98045 98052 98053 98055 98056 98058 98059 98065 98070 98072 98074 98075 98077 98092 98102 98103 98105 98106 98107 98108 98109 98112 98115 98116 98117 98118 98119 98122 98125 98126 98133 98136 98144 98146 98148 98155 98166 98168 98177 98178 98188 98198 98199
zipcode

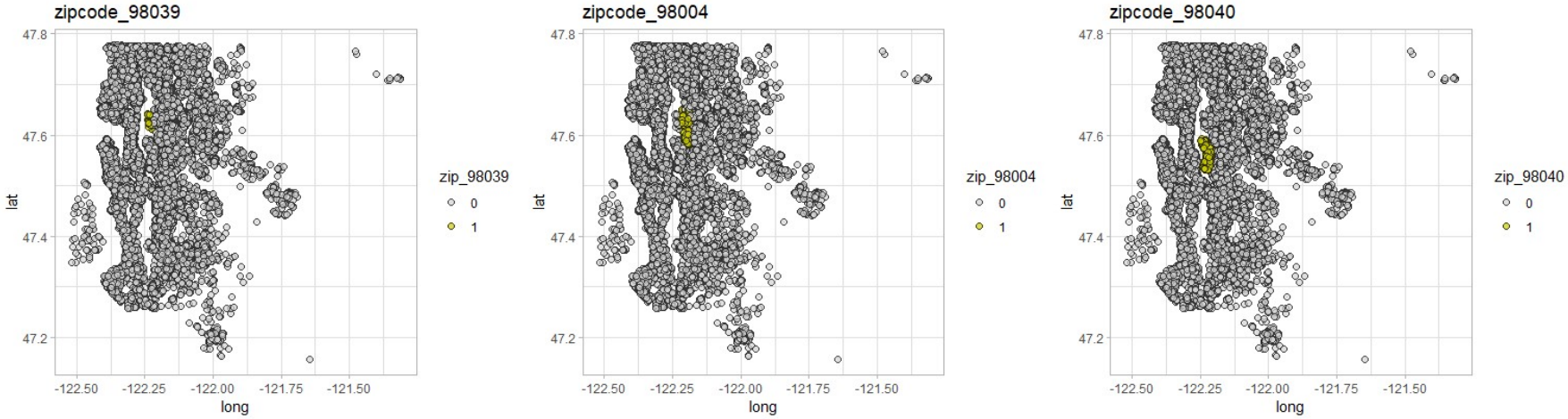

mean log price by zipcode
mean_log price
14.5
14.0
13.5
13.0
12.5
98002 98168 98032 98001 98148 98188 98023 98003 98198 98030 98178 98031 98055 98042 98106 98022 98146 98092 98058 98038 98108 98056 98133 98118 98155 98014 98010 98045 98126 98019 98166 98028 98059 98125 98011 98034 98065 98070 98024 98136 98072 98107 98117 98144 98029 98103 98116 98008 98115 98122 98177 98027 98052 98007 98053 98074 98077 98033 98199 98075 98006 98005 98105 98102 98109 98119 98112 98040 98004 98039
zipcode

lat, long

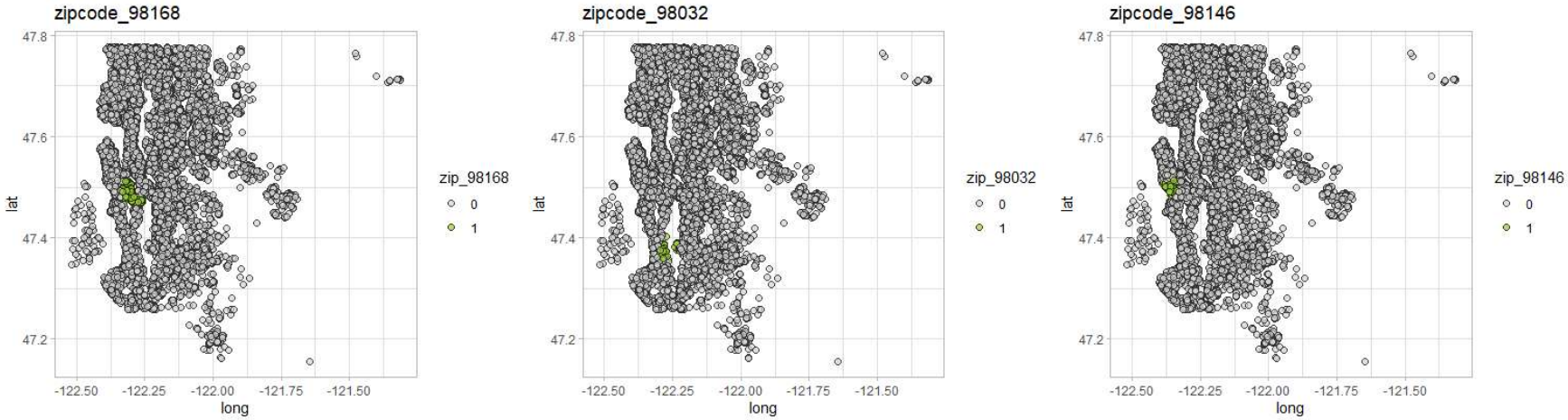

47.8
47.6
47.4
47.2
lat
-122.50
-122.25
-122.00
-121.75
-121.50
long
log(house.train$price)
15
14
13
12
factor(zipcode)
98001 98030 98072 98122
98002 98031 98074 98125
98003 98032 98075 98126
98004 98033 98077 98133
98005 98034 98092 98136
98006 98038 98102 98144
98007 98039 98103 98146
98008 98040 98105 98148
98010 98042 98106 98155
98011 98045 98107 98166
98014 98052 98108 98168
98019 98053 98109 98177
98022 98055 98112 98178
98023 98056 98115 98188
98024 98058 98116 98198
98027 98059 98117 98199
98028 98065 98118
98029 98070 98119

lat, long & zipcode

- 주택 가격이 높은 지역



- 주택 가격이 낮은 지역



lat, long & zipcode

워싱턴주 20대 부촌은 어디?

워싱턴주 최고의 부촌인 메다이나 주민들의 평균 소득은 85만9,379달러에 달한다.

세금시즌이 도래한 가운데 국세청(IRS)에 보고한 소득을 기준으로한 워싱턴주 20대 부촌을 소개한다.

가장 최근 자료인 2016년 IRS 자료를 바탕으로 워싱턴주 톱20개 부촌 우편번호와 지역의 평균 소득이 공개됐다. 모두 킹카운티 내인 이들 지역의 평균 소득은 13만달러가 넘는다.

상위 1,2위 지역은 예상대로 머서 아일랜드와 메다이나이다. 특히 빌 게이츠가 사는 1위 지역 메다이나(98039) 주민들의 평균 수입은 무려 859,379달러로 2위 머서 아일랜드의 2배가 넘는다.

http://www.joyseattle.com/news/37213

▲3위 98004

평균 수입: $266,467

포함 지역: 벨뷰, 클리이드 힐, 부 아츠 빌리지, 야로우 포인트, 헌츠 포인트

▲2위 98040

평균 수입: $314,433

포함 지역: 머서 아일랜드

▲1위 98039

평균 수입: $859,379

포함 지역: 메다이나

5. 데이터 전처리

*** 변수 변환**

- 연속형 데이터 log 변환 : price, sqft_living, sqft_lot, sqft_above, sqft_basement, sqft_living15, sqft_lot15

- 가변수 생성 : zipcode

- 새로운 변수 생성 : yr_renovated 와 yr_built 변수를 합쳐 새로운 **yr_new** 변수를 생성함.

ex) yr_renovated : 0 , yr_built : 1955 → yr_new : 1955

yr_renovated : 2002, yr_built : 1930 → yr_new : 2002

*** 이상치 제거**

- bathrooms가 6개 이상인 것 중에

```
house.train[house.train$bathrooms >= 6,]
```

癤뀀d	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above
947	20150413	5300000	6	6.00	7390	24829	2.0	1	4	4	12	5000
2859	20141007	800000	7	6.75	7480	41664	2.0	0	2	3	11	5080
5108	20141013	7700000	6	8.00	12050	27600	2.5	0	3	4	13	8570
5990	20140811	450000	9	7.50	4050	6504	2.0	0	0	3	7	4050
6469	20140919	6885000	6	7.75	9890	31374	2.0	0	4	3	13	8860
8912	20140505	2280000	7	8.00	13540	307752	3.0	0	4	3	12	9410
10152	20140611	2888000	5	6.25	8670	64033	2.0	0	4	3	13	6120

- bedrooms가 10개 이상인 것 중에

```
house.train[house.train$bedrooms >= 10,]
```

癤뀀d	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above
9280	20140814	1148000	10	5.25	4590	10920	1	0	2	3	9	2500
10575	20141029	650000	10	2.00	3610	11914	2	0	0	4	7	3010

6. Research Questions와 분석 계획

Research Questions

- 주택 가격에 가장 많은 영향을 주는 요인 무엇인가
- 재건축된 주택이 가격에 영향을 주는가
- 면적이 넓고 방이 많고 층이 높으면 가격이 높은가
- 비슷한 면적의 집을 비교하면 어떤 요인이 가격을 결정짓는가
- 어떤 분석방법이 데이터에 가장 적절한가
- 목표 변수를 잘 예측할 수 있는가(test 데이터로 평가)
- 가격이 가장 높은/낮은10개 집의 특징은 무엇인가
- 집의 등급은 지어진 년도와 관련이 있는가 (상관계수 = 0.51)
- 변수들 간의 상관관계는 어떻게 나타나는가
- 우리나라의 강남/강북과 같이 지역에 따라 주택 가격이 높게/낮게 형성되는가
- 가격이 비싼 집의 이웃집도 가격이 높게 나타날 확률이 높은가

분석 계획

- 데이터 표준화(normalization)
- K-means 알고리즘 등으로 이웃 군집화
- 다중회귀분석 (Ridge, Lasso 등)
- 베이지안 기법을 이용한 추정, 검정