Predict house price dataset

Analysis Report

Team member : مروان محمد بدير عبد الله 20160411

Team member : محمود عادل عباس 20160394

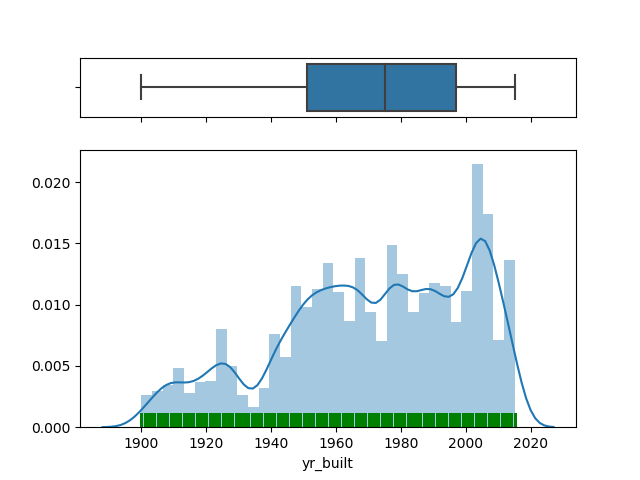
|  |  |  |
| --- | --- | --- |
| **attribute** | **mean** | **variance** |
| id | **4580301520.9** | **8.3 e+18** |
| price | **540088.1** | **134782378397.1** |
| bedrooms | **3.4** | **0.9** |
| bathrooms | **2.1** | **0.6** |
| sqft\_living | **2079.9** | **843533.7** |
| sqft\_lot | **15107** | **1715658774.2** |
| floors | **1.5** | **0.3** |
| waterfront | **0.0075** | **0.0075** |
| view | **0.2** | **0.6** |
| condition | **3.4** | **0.4** |
| grade | **7.7** | **1.4** |
| sqft\_above | **1788.4** | **685734.7** |
| sqft\_basement | **291.5** | **195872.6** |
| yr\_built | **1971** | **862.8** |
| yr\_renovated | **84.4** | **161346.2** |
| lat | **47.5** | **0.019** |
| sqft\_living15 | **1986.5** | **469761.2** |
| sqft\_lot15 | **12768.5** | **745518225.3** |
| long | **-122.2** | **0.019** |
| Zip code | **98078** | **2862.8** |

**Year Built Attribute**

This figure represents the density of buildings each year and show that all buildings are built in 20s

We notice that more buildings are built in recent years from ( the mean value that equal 1971.0051357978994 )

The variance value = 862.7972621659717

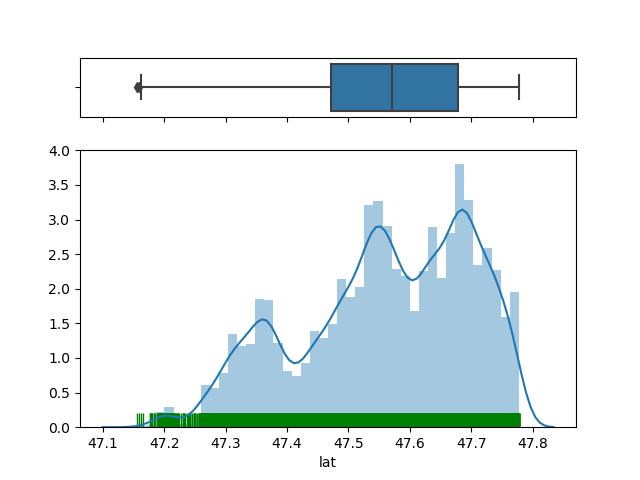


**Local area transport attribute**

This figure represents the density of buildings according to LAT values that's range is between (47.1 : 47.8 )

The mean value = 47.56005251931704 we notice from the mean value that most buildings has good transportation

The variance value = 0.01919990179600804

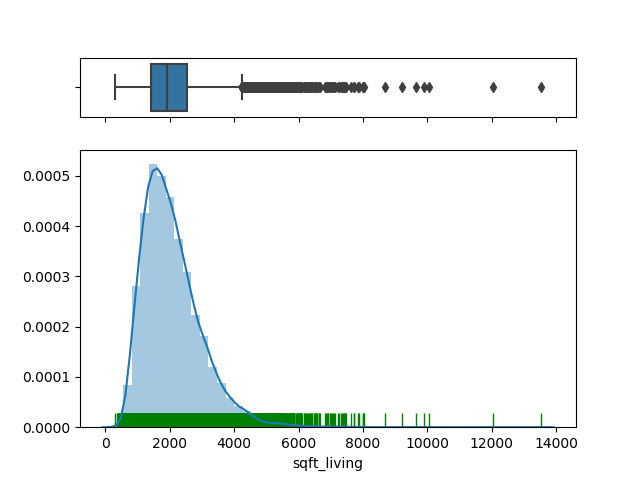


**Square feet of living attribute**

This figure represents the density of buildings according to sqft\_living attribute

we notice from this value of mean = 2079.8997362698374 that most buildings has sqft\_living between the small ranges and

we notice from the variance value = 843533.6813681519 that the distribution has a big spread and has many outliers



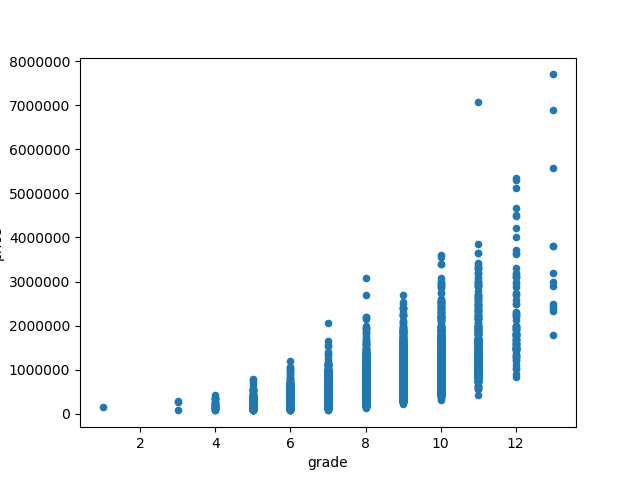
**Scatter plot**

This figure represents the scatter plot between (Grade(overall grade given to the housing unit, based on King County grading system) in X axis and Price in Y axis)

We notice the positive relationship (the price become high when the value of grade is high )

And if we calculate the covariance between them we will find it equal to 288026.5 which means that there is a positive relationship between them also

And that is true by logic because it’s normal when price increase if the grade for this house is high



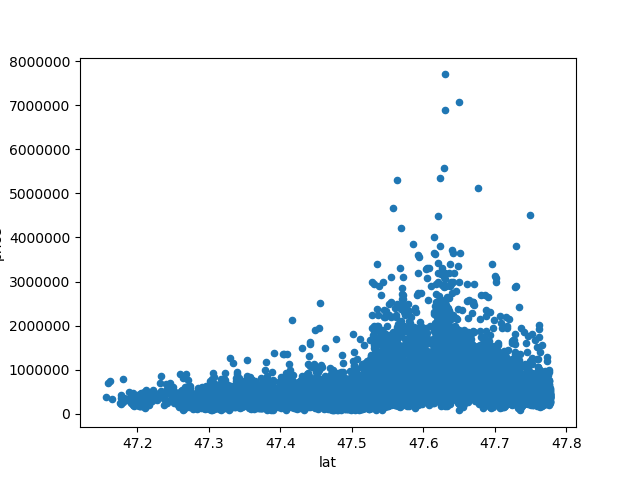
**Scatter plot 2**

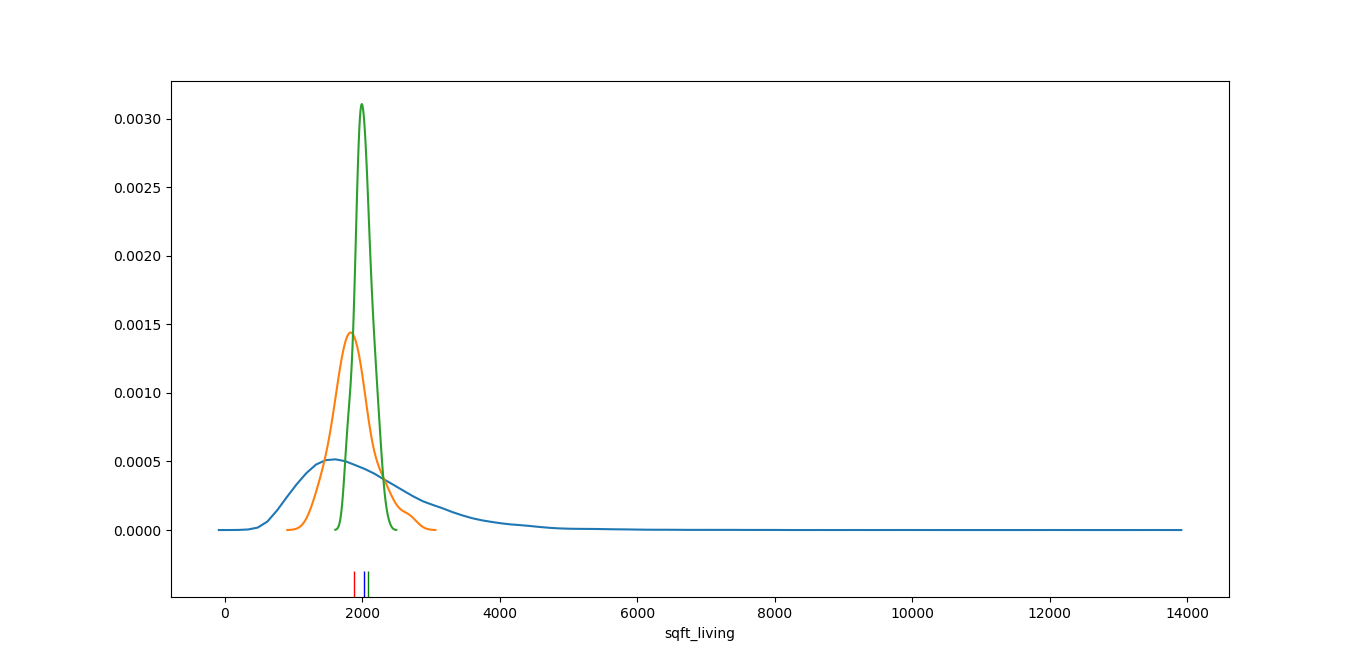
This figure represents the scatter plot between (LAT in X axis and Price in Y axis)

We notice the positive relationship (the price become high when the value of LAT is high )

And if we calculate the covariance between them we will find it equal to 15617.3 which means that there is a positive relationship between them also

And that is true by logic because it’s normal when price increase if the local transport for this house is high but it’s decrease when it’s very high that’s might because it became very noisy place if there are many local transport for it (that’s our estimation)

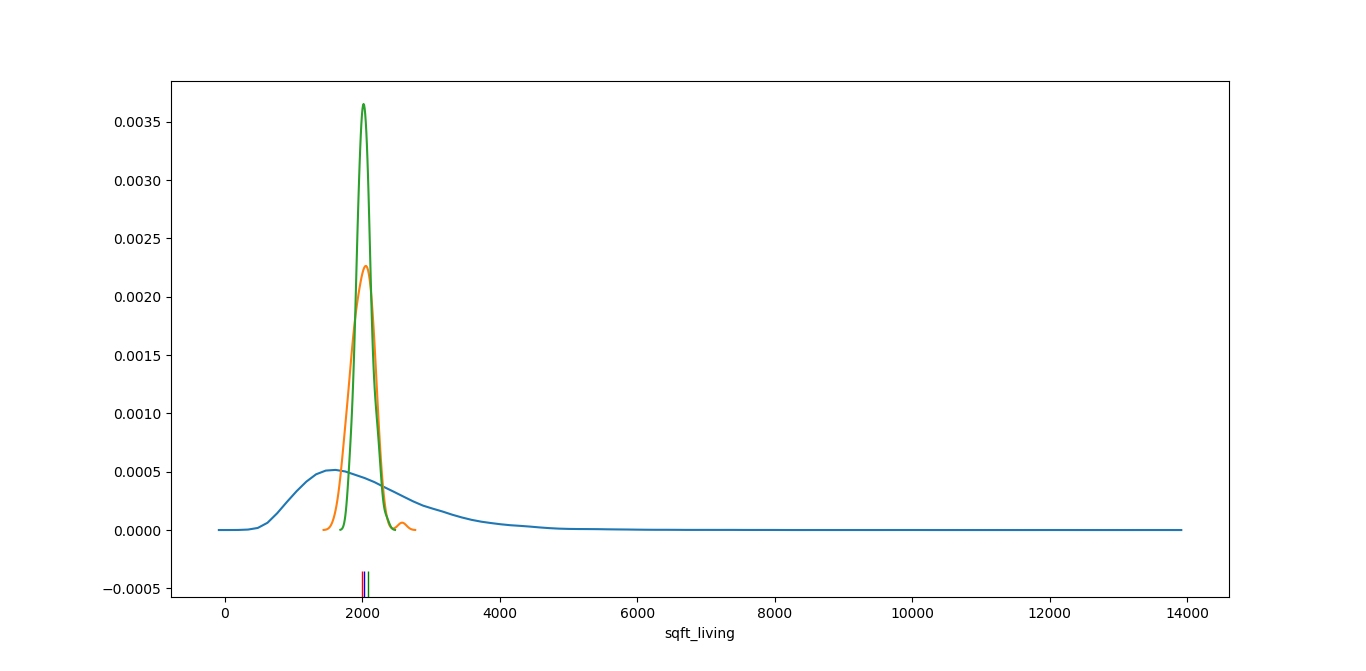




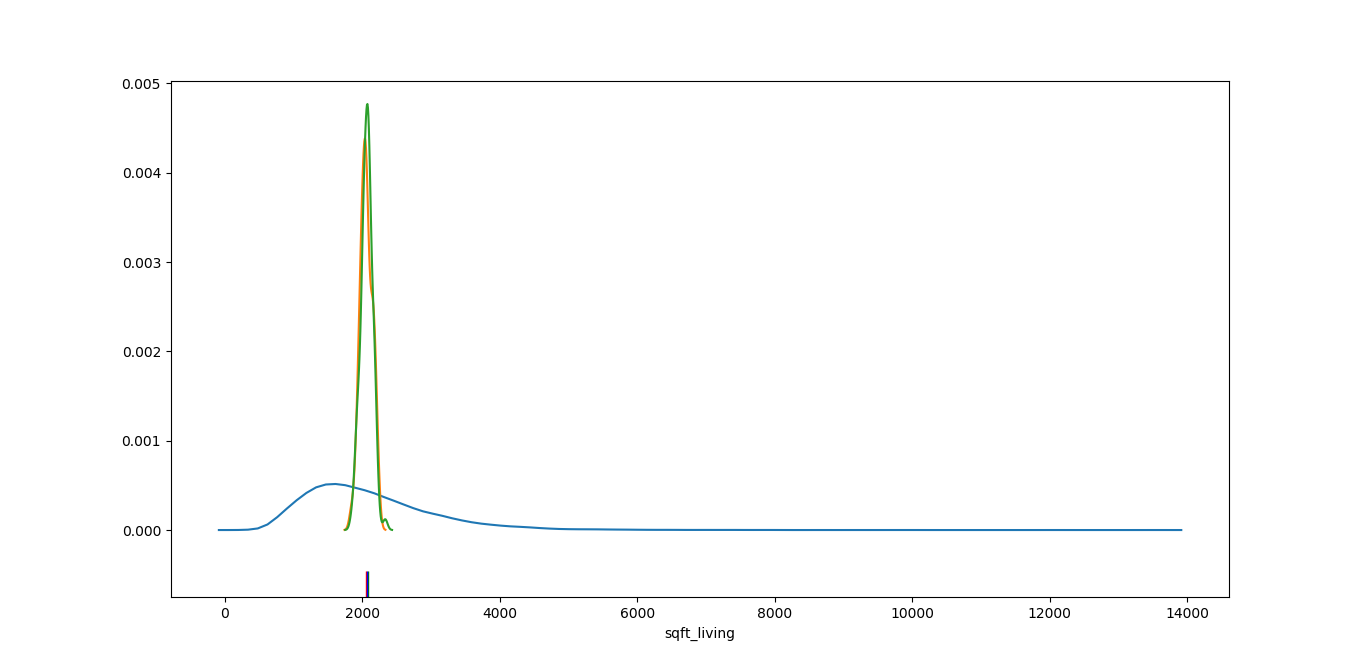
This figure represents the PDF for sqft\_living (the green one) and plot it’s means and PDF for samples means if I take 100 samples randomly from sqft\_living with sample size 10 (the red one)

and PDF for samples means if I take 100 samples randomly from sqft\_living with sample size 50 (the blue one) and there mean of samples mean.

And here we can see that the PDF’s are different but the mean of samples mean is close to the original mean (the green one) and also we can see the blue mean of samples is closer to tha original mean than the red one , So what will be happened if we increase samples size ? we will know that at the next figure



This figure represents the same thing but we increase the sample size for the red one to 50 and the blue one to 100 with the same number of samples (100) what is happened here is that the mean of samples means for both PDF’s are move closer to the original mean which mean that the more we increase sample size the more we get closer to the original mean



Here we increase the sample size for the red one to 90 and the blue one to 120 with the same number of samples (100)

And this is a covariance matrix for 5 attributes in our dataset

price sqft\_living sqft\_above sqft\_basement grade

price 1.347824e+11 2.367154e+08 1.841014e+08 5.261406e+07 288026.499738

sqft\_living 2.367154e+08 8.435337e+05 6.666978e+05 1.768358e+05 823.407664

sqft\_above 1.841014e+08 6.666978e+05 6.857347e+05 -1.903683e+04 735.805403

sqft\_basement 5.261406e+07 1.768358e+05 -1.903683e+04 1.958727e+05 87.602262

grade 2.880265e+05 8.234077e+02 7.358054e+02 8.760226e+01 1.381703