Northeastern University
College *of* Engineering

IE 6200

PROBABILITY AND STATISTICS

Spring 2019

CALIFORNIA HOUSING DATA REPORT

TEAM MEMBERS

PAVAN KUMAR REDDY

DILIP PUROHIT

LING ZHAO

OBJECTIVES

- Cleaning the dataset to make sure they are no null values in it.
- Understanding the dataset and analysing every column and their relations with each other.
- Getting the visualisations for every attribute in the dataset.
- Performing Sampling for a column to perceive the difference between sample and population
- Executing some of the probability and statistical related analysis to get an overview of the data.
- Identifying the distribution of attributes and obtain the confidence intervals for population.
- Carrying Hypothesis testing procedure to evaluate which hypothesis is best supported by the sample data.

DATA DESCRIPTION

Source:

This dataset is a modified version of the California Housing dataset available in kaggle.The dataset may also be downloaded from StatLib mirrors.

This dataset appeared in a 1997 paper titled *Sparse Spatial Autoregressions* by Pace, R. Kelley and Ronald Barry, published in the *Statistics and Probability Letters* journal. They built it using the 1990 California census data. It contains one row per census block group. A block group is the smallest geographical unit for which the U.S. Census Bureau publishes sample data (a block group typically has a population of 600 to 3,000 people).

Our data set has the following columns:

Longitude	:	Continuous Data
Latitude	:	Continuous Data
HousingMedian_Age	:	Continuous Data
Total_Rooms	:	Continuous Data
Total_Bedrooms	:	Continuous Data
Population	:	Continuous Data
Households	:	Continuous Data
Median_Income	:	Continuous Data
Median_House_Value	:	Continuous Data

- **Latitude** and **Longitude** which are used to specify the particular location.
- **HousingMedianAge**: Median age of the house with in a block. A lower number specifies the newer building.
- **TotalRooms**: Total number of rooms within a block.
- **TotalBedRooms**: Total number of bed rooms within a block.
- **Population**: Total number of people residing in a particular block.
- **Households**: Total number of households for a block.
- **MedianIncome**: MedianIncome per month for the households within a blocks of houses(measured in thousands of US dollars)
- **MedianHouseValue**: Median house value for households with in a block(measured in tens US dollars)

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income
1	-122.22	37.86	21	7099	1106	2401	1138	8.301
2	-122.24	37.85	52	1467	190	496	177	7.257
3	-122.25	37.85	52	1274	235	558	219	5.643
4	-122.25	37.85	52	1627	280	565	259	3.846
5	-122.25	37.85	52	919	213	413	193	4.036
6	-122.25	37.84	52	2535	489	1094	514	3.659
7	-122.25	37.84	52	3104	687	1157	647	3.120
8	-122.26	37.84	42	2555	665	1206	595	2.080
9	-122.25	37.84	52	3549	707	1551	714	3.691
10	-122.26	37.85	52	2202	434	910	402	3.203

Showing 1 to 11 of 20,639 entries

```
summary(housing_df)
  longitude         latitude     housing_median_age  total_rooms     total_bedrooms
Min.   :-124.3   Min.   :32.54   Min.   : 1.00      Min.   :    2   Min.   :   1.0
1st Qu.:-121.8   1st Qu.:33.93   1st Qu.:18.00      1st Qu.: 1448   1st Qu.: 295.5
Median :-118.5   Median :34.26   Median :29.00      Median : 2127   Median : 435.0
Mean   :-119.6   Mean   :35.63   Mean   :28.64      Mean   : 2636   Mean   : 537.9
3rd Qu.:-118.0   3rd Qu.:37.71   3rd Qu.:37.00      3rd Qu.: 3148   3rd Qu.: 647.0
Max.   :-114.3   Max.   :41.95   Max.   :52.00      Max.   :39320   Max.   :6445.0
  population      households      median_income     median_house_value
Min.   :    3   Min.   :   1.0   Min.   : 0.4999   Min.   : 14999
1st Qu.:  787   1st Qu.: 280.0   1st Qu.: 2.5631   1st Qu.:119600
Median : 1166   Median : 409.0   Median : 3.5347   Median :179700
Mean   : 1426   Mean   : 499.6   Mean   : 3.8705   Mean   :206844
3rd Qu.: 1725   3rd Qu.: 605.0   3rd Qu.: 4.7428   3rd Qu.:264700
Max.   :35682   Max.   :6082.0   Max.   :15.0001   Max.   :500001
```

ASSUMPTIONS

As our data is huge,we assumed the dataset as population data.

Checking null values:

Firstly, after importing the dataset ,we made sure that there are no null values which would create errors unless they are made to 0 or as median of their respective columns.

```
checking_null<-is.na(housing)
View(checking_null)
```

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income
1	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
2	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
3	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
4	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
5	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
6	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
7	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
8	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
9	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
10	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

Here we can see FALSE which means there are no null values present in the dataset.

Later correlation coefficient is calculated between every two attributes to understand the extent of their relationships.

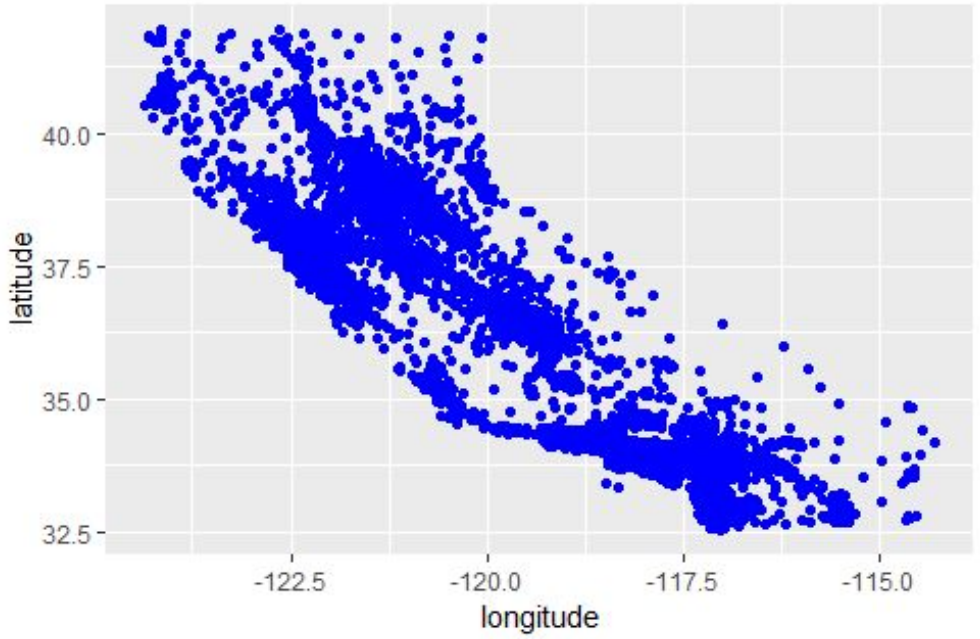
```
View(cor(housing_df)
```

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value
longitude	1.00000000	-0.92466103	-0.10814077	0.04451880	0.068319557	0.099717085	0.05525087	-0.015027656	-0.045836
latitude	-0.92466103	1.00000000	0.01112314	-0.03606009	-0.066272188	-0.108740473	-0.07098915	-0.079941500	-0.144288
ousing_median_age	-0.10814077	0.01112314	1.00000000	-0.36123801	-0.320453318	-0.296211606	-0.30288360	-0.119164227	0.105536
total_rooms	0.04451880	-0.03606009	-0.36123801	1.00000000	0.929891314	0.857121141	0.91848205	0.198170585	0.134253
total_bedrooms	0.06831956	-0.06627219	-0.32045332	0.92989131	1.000000000	0.878019990	0.97982804	-0.007984224	0.050700
population	0.09971708	-0.10874047	-0.29621161	0.85712114	0.878019990	1.000000000	0.90721799	0.004945822	-0.024552
households	0.05525087	-0.07098915	-0.30288360	0.91848205	0.979828035	0.907217986	1.00000000	0.013146107	0.065952
median_income	-0.01502766	-0.07994150	-0.11916423	0.19817059	-0.007984224	0.004945822	0.01314611	1.000000000	0.688000
edian_house_value	-0.04583660	-0.14428861	0.10553613	0.13425302	0.050700679	-0.024552395	0.06595224	0.688000493	1.000000

Based on the correlations we can discern that the correlation coefficient values which lie between 0.8 to 1 or -0.8 to -1 have very strong correlations and 0.6 to 0.8 have strong, 0.4 to 0.6 have weak and lower that will have weaker correlations.

From the table we get 0.688 correlation coefficient value between median_income and median_house_value which is strong and are relatable. Similarly for total rooms and total bedrooms , have very strong correlation of 0.93.

For getting the exact locations we plot a graph for longitudes and latitudes. Every point in the graph below is a location.

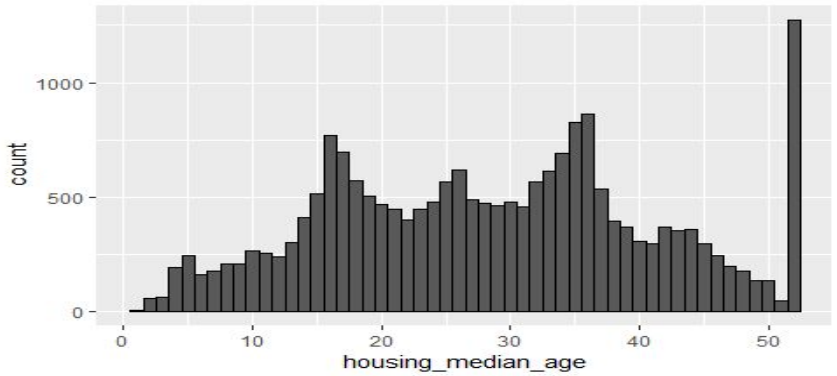


Here we can see 20,639 locations distributed on the graph.

For understanding the data we performed some visualisations like histogram, line graph, scatter plot and other inferential plots.

Histogram for median_age of the house

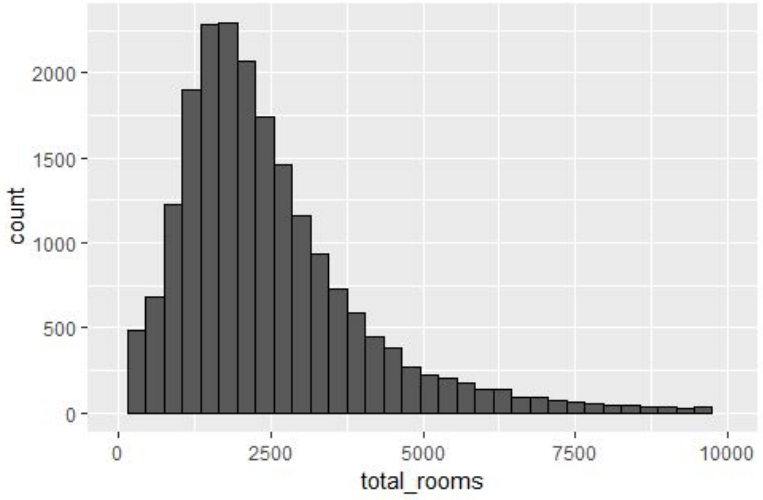
```
ggplot(data=housing_df)+
  geom_histogram(mapping = aes(x=housing_median_age),binwidth=1,color='black')
```



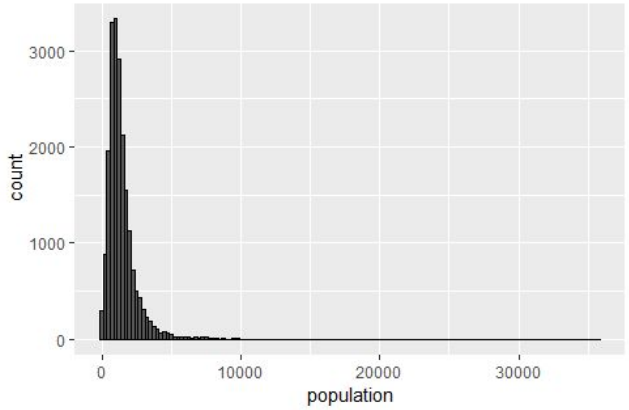
From the above histogram we see that most of the people live in the houses that are of median ages between 15 to 35 years.

Histogram for total_rooms

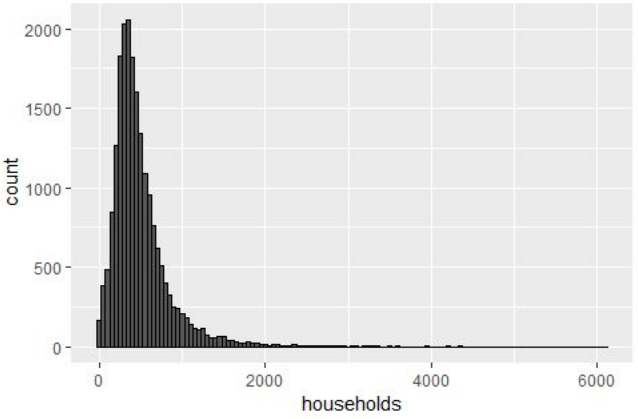
Most of the people are living in a locality with a total rooms of over 2000 and a very few of them live in a locality of around 10000 rooms. It is a right skewed distribution.



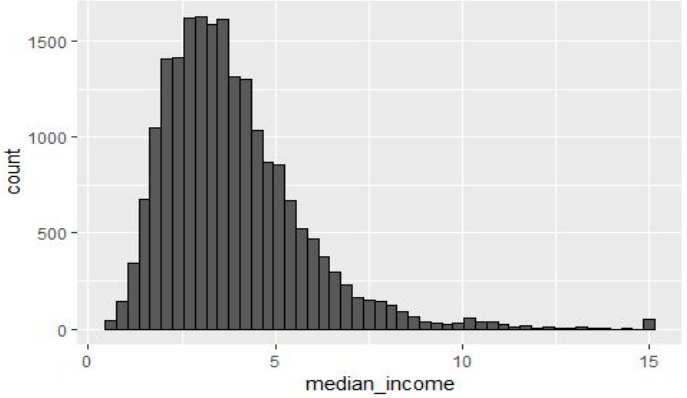
Histogram for population



Histogram for Households

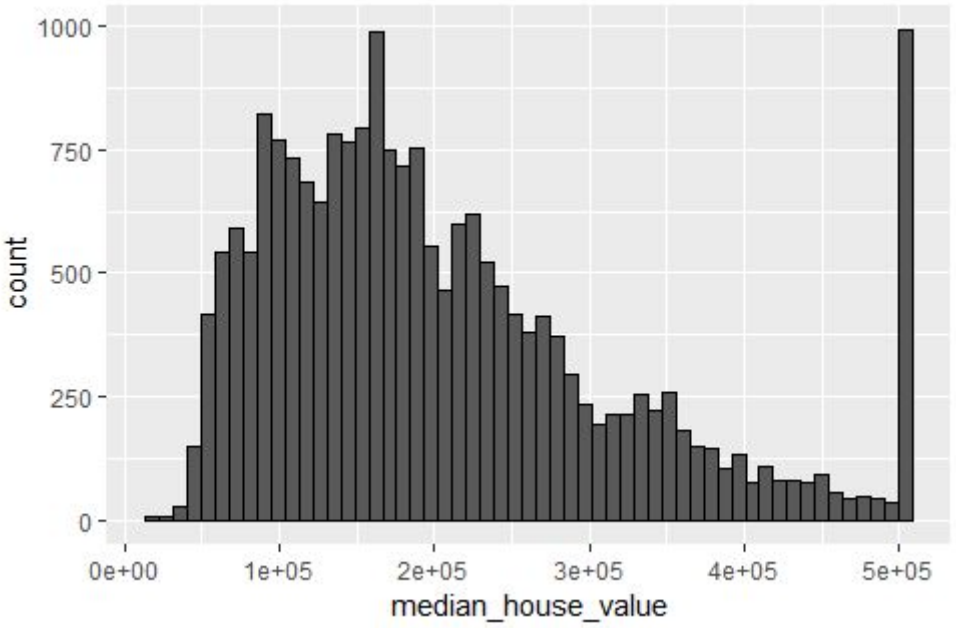


Histogram for median Income of the population



Most of the people have a median income around 5000$ per month and a very few of them get over 10000$ per month. We can see that it is right skewed distribution.

Histogram for median_house_value



Most of the people live in house worth between 1M$-2M$. And it looks like a right skewed distribution.

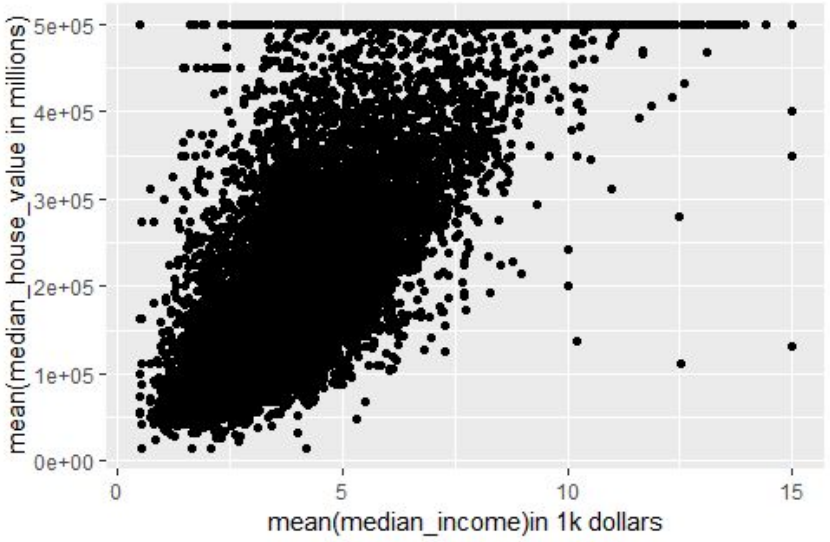
In the data set we have same latitudes and longitudes repeated . So we grouped according to latitudes and longitudes and for the other columns of same latitudes and longitudes, mean is considered.

```
grouping<-group_by(housing_df,longitude,latitude)
View(grouping)
grouped_mean<-aggregate(grouping[,3:9],by=list(housing_df$longitude,housing_df$latitude),FUN=mean)
View(grouped_mean)
```

	Group.1	Group.2	housing_median_age	total_rooms	total_bedrooms	population	households	medi
1	-117.04	32.54	7.00000	938.0000	297.0000	1187.0000	282.0000	
2	-117.09	32.55	8.00000	6533.0000	1217.0000	4797.0000	1177.0000	
3	-117.06	32.55	5.00000	3223.0000	940.0000	3284.0000	854.0000	
4	-117.04	32.55	15.00000	2206.0000	648.0000	2511.0000	648.0000	
5	-117.12	32.56	20.00000	2524.0000	682.0000	1819.0000	560.0000	
6	-117.10	32.56	16.00000	2687.0000	501.0000	1502.0000	480.0000	
7	-117.09	32.56	8.00000	864.0000	156.0000	626.0000	172.0000	
8	-117.07	32.56	9.00000	3648.0000	895.0000	3293.0000	840.0000	
9	-117.06	32.56	11.00000	2754.5000	804.0000	2958.0000	765.5000	
10	-117.05	32.56	19.00000	1457.3333	372.0000	1351.6667	365.3333	

Showing 1 to 11 of 12,589 entries

Now we plotted to median_income against median_house_value with respect to every grouped latitude and longitude.

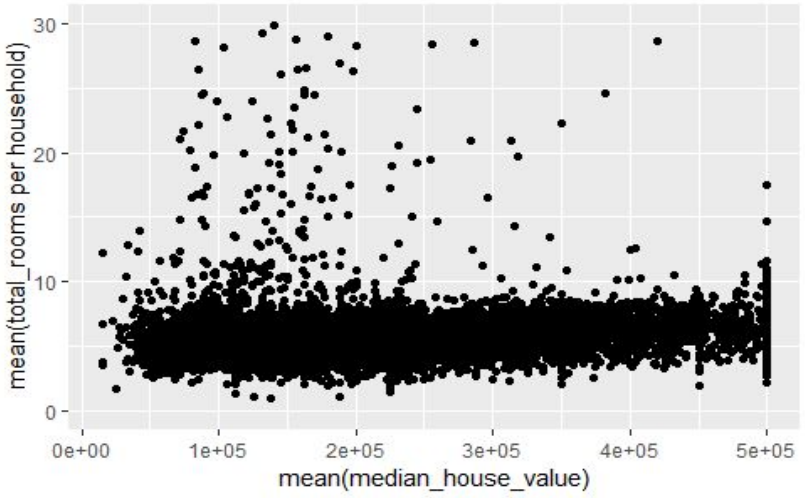


From this plot we can understand that maximum population have less than 7000$ per month in our dataset and they mostly live in the houses ranging from 1M$-4M$ as it is shown very dense in the region. People who earn around 10k$ per month live in houses worth over 4M$ and a few of them live below that figure.

Graph between median_house_value and total_rooms per household

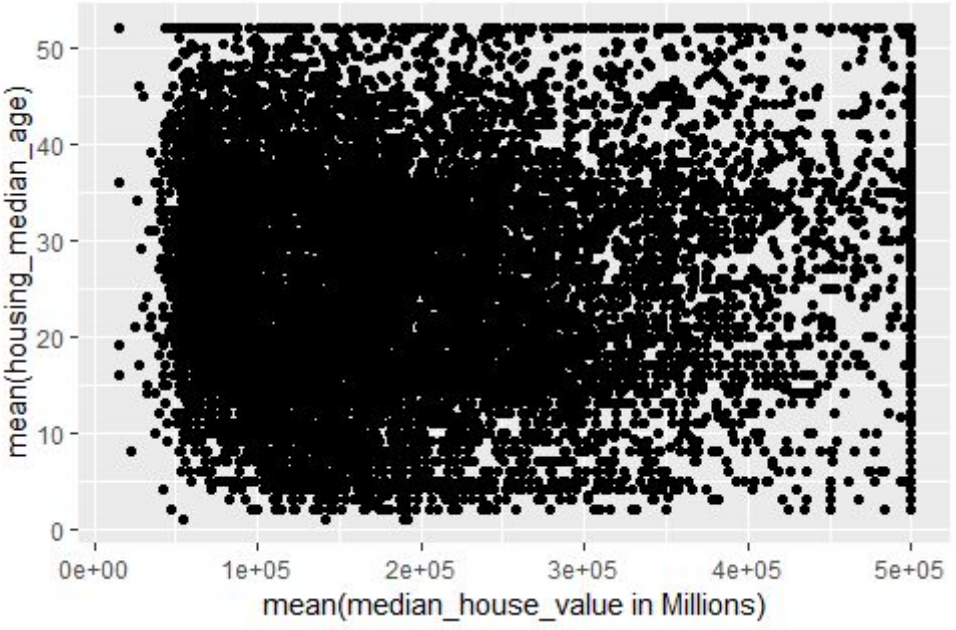
To get total_rooms per household we have divided total_rooms in the locality by number of household in the locality.

```
ggplot(data=grouped_mean)+
  geom_point(mapping =
aes(x=median_house_value,y=(total_rooms/households)),color='black')+
  xlab("mean(median_house_value)")+
  ylab("mean(total_rooms)")+ylim(0,30)
```



From the above scatter plot that on an average the total number of room per household is mostly below 8. And there are all types of houses ranging from 1M$-5M$ with that number of rooms. The total rooms of certain household have around 10-30 and are mostly less than 2M$.

Graph between Household value and median_house_value



From this graph we can understand that age doesn't matter much to the people. Most of the people live even in 50 year old houses for the house value less than 2.5M$ and after that it is rarely chosen.

Sampling:

To compare the differences between sample and population we have taken a random sample size of 100 from **median_house_age**.

Population

Mean of housing_median_age = 28.63889

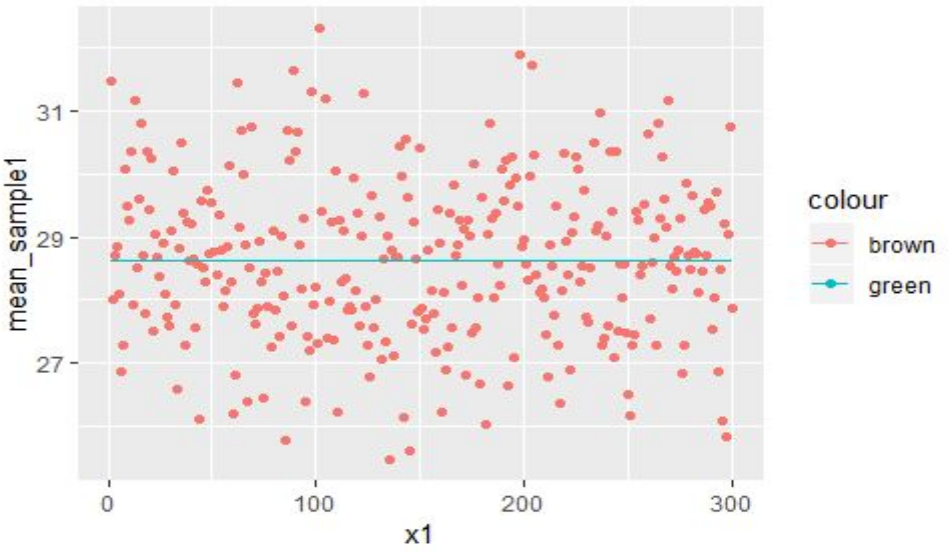
Standard deviation = 12.58557

From the **sample** of 100, we got

Mean = 28.63387

Standard deviation = 1.284474

We have got almost the same mean values for both sample and population data.



Here the green line is the population mean for the **median_house_age** and the red points are the distributed means of the sample data.

Now we tried for **median_house value** to identify the difference between sample and population data.

Population

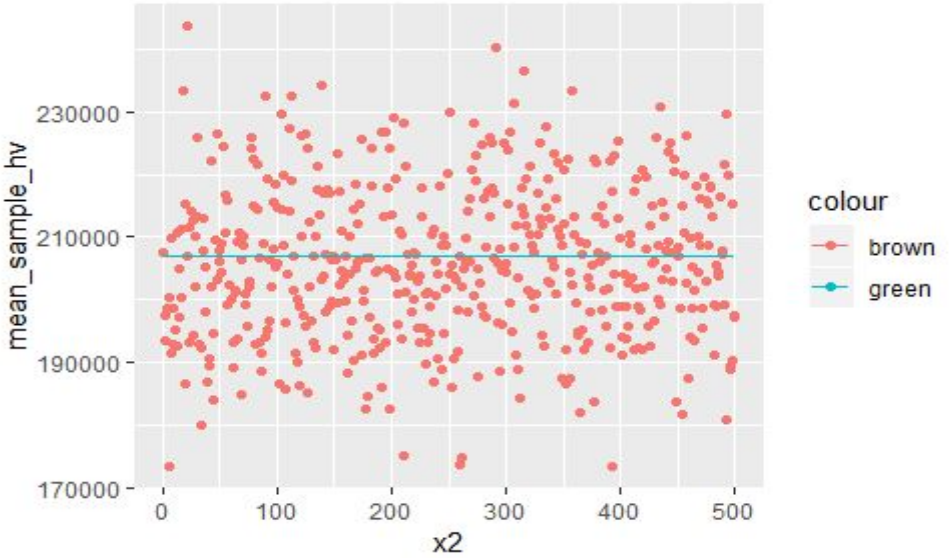
Mean = 206843.9

Variance = 13313867080

Sample data of 100

Mean = 205813.5

Variance = 145421770



Here is the plot for the means of sample and population data of **median_housing_value.**

The green line is population mean and the sample means are spread out in brown color. And both the means are almost same.

Goodness Of Fit tests:

We are selecting the suitable distribution for each and every column in our dataset. For this we used the fitdistrplus package to find the correct distribution with respect to the graph. We also used descdist() function to compute the descriptive statistics.

For example we are showing the code for hosuing_median_age column and for all other columns we will just upload the graph and the necessary plots

DESCRIPTIVE STATISTICS AND GOODNESS OF FIT GRAPH FOR HOUSING_MEDIAN_AGE:

```
#DESCRITIVE STATISTICS
ds3<-descdist(housing_df$housing_median_age)
ds3
fit3_nor<-fitdist(housing_df$housing_median_age, "norm")
summary(fit3_nor)
```

OUTPUT

summary statistics

---

min: 1 max: 52

median: 29

mean: 28.63889

estimated sd: 12.58557

estimated skewness: 0.06043027

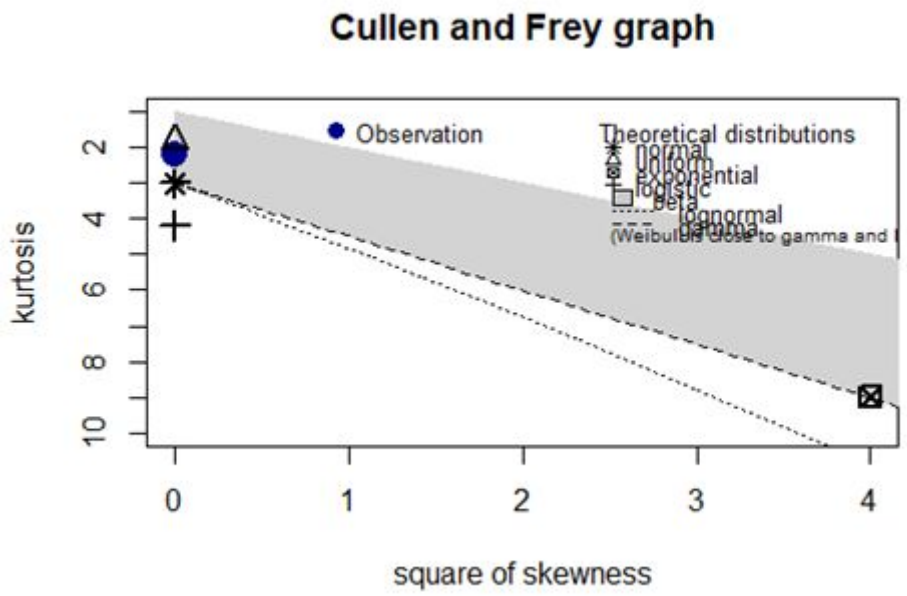
estimated kurtosis: 2.199437

From the above statistics we can conclude that skewness=0.06 and kurtosis=2.199

Generally, for the perfect normal distribution we can say that the skewness=0 and the kurtosis =3.

By using the skewness and kurtosis we can determine we type of distribution we have.

CULLEN AND FREY GRAPH:



Conclusion:

From the above graph we can observe that the observation(blue dot) is between normal and uniform distribution.

To check for the perfect distribution between these uniform and normal distribution we will compare the parameters of the fitting distribution. After considering those parameters we will decide which one is best. In the above case normal distribution is the perfect distribution.

OUTPUT:

Fitting of the distribution ' norm ' by maximum likelihood

Parameters :

estimate Std. Error

mean 28.63889 0.08760280

sd 12.58526 0.06194453

Loglikelihood: -81554.29 AIC: 163112.6 BIC: 163128.4

Correlation matrix:

mean sd

mean 1 0

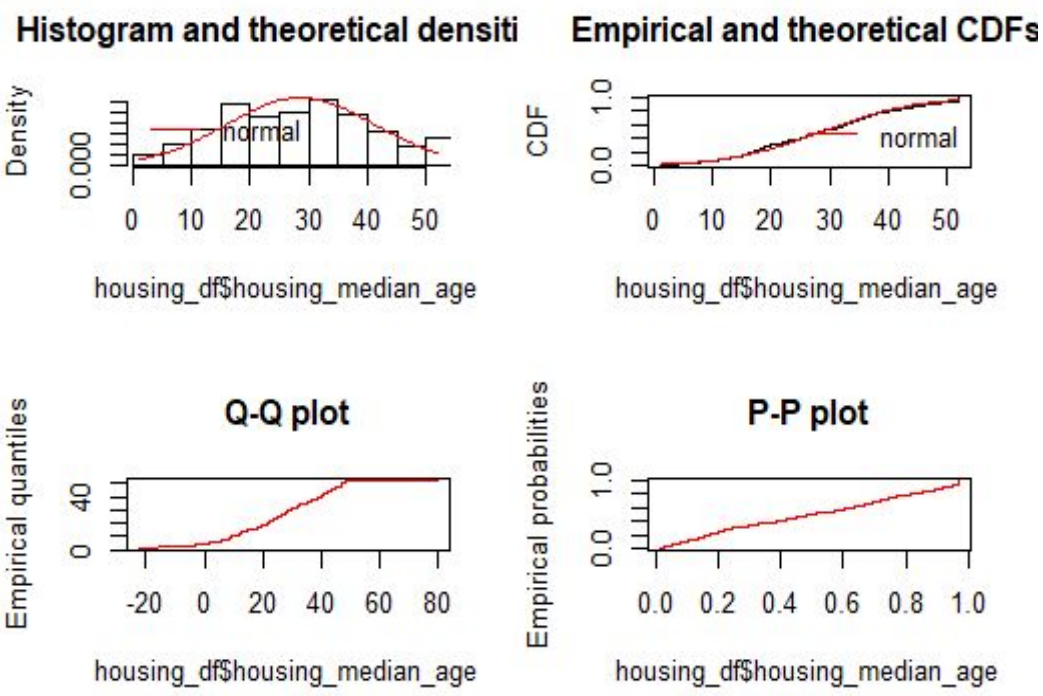
sd 0 1

Goodness Of fit graphs:

Code:

```
par(mfrow=c(2,2))
plot.legend <- c("normal")
denscomp(list(fit3_nor), legendtext = plot.legend, xlab =
'housing_df$housing_median_age', xlegend = 'topleft')
cdfcomp (list(fit3_nor), legendtext = plot.legend, xlab =
'housing_df$housing_median_age')
qqcomp  (list(fit3_nor), legendtext = plot.legend, xlab =
'housing_df$housing_median_age')
ppcomp  (list(fit3_nor), legendtext = plot.legend, xlab =
'housing_df$housing_median_age')
```

OUTPUT:



1) *For total rooms*: This column follows lognormal distribution

Descriptive statistics output:

summary statistics

---

min: 2 max: 39320

median: 2127

mean: 2635.848

estimated sd: 2181.634

estimated skewness: 4.147347

estimated kurtosis: 35.63077

Fitting of distribution parameters:

Fitting of the distribution ' lnorm ' by maximum likelihood

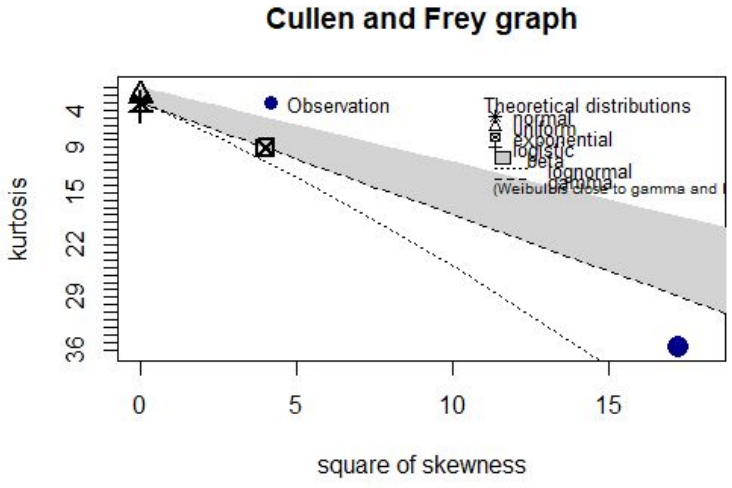
Parameters :

	estimate	Std. Error
meanlog	7.6286183	0.005231089
sdlog	0.7515129	0.003698909

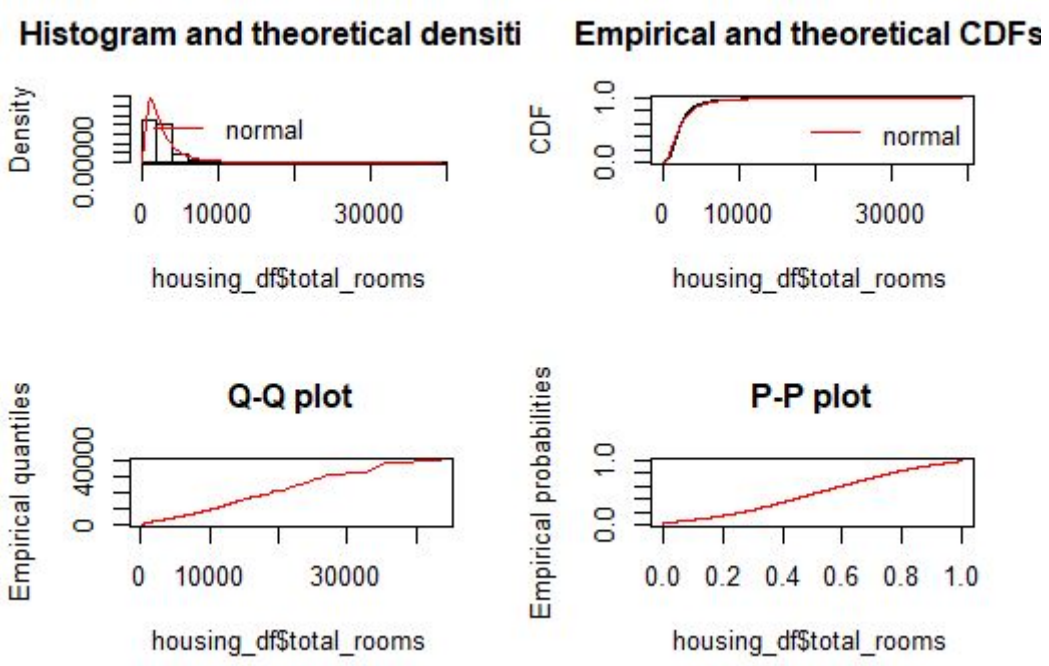
Loglikelihood: -180836.6 AIC: 361677.3 BIC: 361693.2

Correlation matrix:

	meanlog	sdlog
meanlog	1	0
sdlog	0	1



Goodness of fit graph:



2) *Total Bedrooms* : This perfectly fits into lognormal distribution

OUTPUT:

summary statistics

---

min: 1 max: 6445

median: 435

mean: 537.9178

estimated sd: 421.2485

estimated skewness: 3.453129

estimated kurtosis: 24.92387

Fitting parameters:

Fitting of the distribution ' lnorm ' by maximum likelihood

Parameters :

estimate Std. Error

meanlog 6.051092 0.005100019

sdlog 0.732683 0.003606228

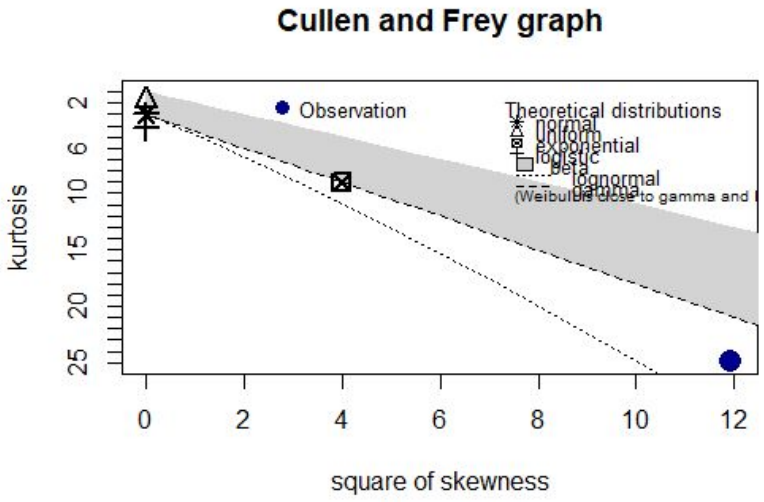
Loglikelihood: -147754.4 AIC: 295512.7 BIC: 295528.6

Correlation matrix:

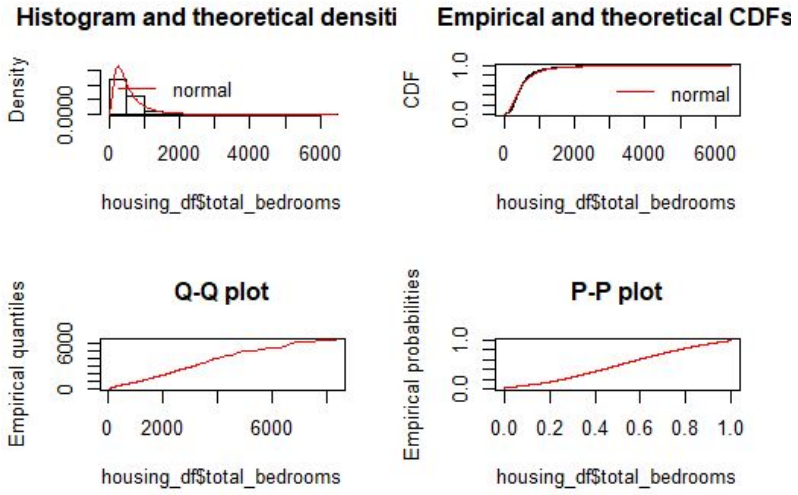
meanlog sdlog

meanlog 1 0

sdlog 0 1



Goodness of fit graph:



Confidence Interval: (For median_house_age)

We considered α which is the confidence interval as 97% and found the range of the population mean.

Code:

```
pop_sd_median_age<-sd(housing_df$housing_median_age)
pop_sd_median_age
sample_size<-c(100)
error<-((qnorm(0.97)*pop_sd_median_age)/sqrt(100))
error
left<-mean_means_age-error
left
right<-mean_means_age+error
right
```

OUTPUT:

pop_sd_median_age

12.58557

error

2.367086

left

26.15451

right

30.88869

Conclusion:

From the above values I can conclude that for the confidence interval of 97% we can

Say that 26.15451< population mean < 30.88869

Hypothesis testing:

One sample z-test is used when the population is normally distributed and the population variance is known. In our case we are using this one sample z test because we know the population variance and our population is normally distributed.

Defining Hypothesis:

Null Hypothesis(H0): $\mu = \mu 0$

We are defining our Null hypothesis as the mean of the population is equal to the mean of the sample.

Alternate Hypothesis(Ha): $\mu != \mu 0$

We are defining Alternate Hypothesis as exactly opposite of the Null Hypothesis i.e mean of the population is not equal to the sample

- In our case we are selecting housing_median _age because it is the only column which is normally distributed.

Step 1: Creating a sample

We already created a sample and we are using the same sample

Step 2: Considering the significance value is alpha=0.01

Step 3: Calculating the value of z

Code:

```
z_test<- (((mean_means_age)-(pop_mean_age))/((pop_sd_age)/(sqrt(100))))
z_test
```

OUTPUT:

z_test

0.01039216

Conclusion:

Since the z-value lies within the range [−1.96, 1.96], we thus fail to reject null hypothesis

conclusion is that there is no significant difference between sample housing_median_age and population housing_median_age.