Selling houses in the most profitable way

Yuqing Zhang- 1005842368

April 18, 2021

```
knitr::opts_chunk$set(warning = FALSE, message = FALSE)
library(openintro)

## Loading required package: airports

## Loading required package: cherryblossom

## Loading required package: usdata

## Warning: replacing previous import 'lifecycle::last_warnings' by
## 'rlang::last_warnings' when loading 'tibble'

## Warning: replacing previous import 'lifecycle::last_warnings' by
## 'rlang::last_warnings' when loading 'pillar'

library(tidyverse)

## Warning: replacing previous import 'lifecycle::last_warnings' by
## 'rlang::last_warnings' when loading 'hms'

## ── Attaching packages ─────────────────────────────────────── tidyverse
1.3.1 ──

## ✓ ggplot2 3.3.3     ✓ purrr   0.3.4
## ✓ tibble  3.1.2     ✓ dplyr   1.0.6
## ✓ tidyr   1.1.3     ✓ stringr 1.4.0
## ✓ readr   1.4.0     ✓ forcats 0.5.1

## ── Conflicts ──────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

Abstract

This report is dedicated to analyze the relationship between houses price and houses age so that people can have some idea for the house market in King County, USA. And then further get a general idea on the global house market trend. There are 5 main sections in my report. They are introduction, data, methods, results and conclusions.
The report introduces the background information for house market globally and then turn the audience’s attention to the relationship between houses age and houses price in the

introduction section. Also, we give a hypothesis and roughly view about methods in this section.

When the report moves to the data part. In this section, I show the dataset which is used in the report and other relevant numerical summaries such as mean, proportion, etc. and graphs such as histograms, barplots, etc. So that the auidance get a general idea on the information which is about houses age and houses price in King County, USA

It is obvious that using methods to approach our aim is important. We introduce the definition and usage of six methods which are maximum likelihood estimator derivation, confidence interval via empirical bootstrap, hypothesis test of the mean, goodness of fit test, Bayesian credible interval and simple linear regression modal in the Methods section. Methods following this order.

By using methods, we get the results. We first introduce the overall results in the Results section then analysis the results that approach by each methods. The report put results in table and all calculations can be found in Appendix.

Lastly, by drawing a conclusion, the report gives the relationship and detailed information for houses price and houses age in King County. Then give advice and general estimate for the global market.

Introduction

The real estate market has been receiving widespread attention. People have need to buy houses and also have need to sell houses. Nowadays, housing prices all over the world are rising year by year. This is the result of vertical comparison. However, comparing through horizontal is also very important. That means there are also many factors that affect houses price, such as age under the same time period. Inspired by this, the topic of the report is the best choice for selling a house. Through analyzing various indicators of houses age and houses price, the report shows the relationship between them. And then further demonstrate any general ideas and views for the global market. In this report, we assume that “New houses have higher price and selling houses when it is new will give more profit to the sellers.” The analysis in this report is important since the ordinary people may have limited consideration for selling houses. They may lose money or even fail to sell. Hence, giving the analysis is pointing a way out for them.

The report chooses to use a dataset which is called “House Sales in King County, USA”. The data is downloaded from https://www.kaggle.com/harlfoxem/housesalesprediction. Data in a dataset that is relevant with the task of the report will be chosen to use. Hence, the report will focus on King County, USA and further give advice globally.

Moreover, graphs combines with methods which are maximum likelihood estimator derivation(MLE), confidence interval via empirical bootstrap, hypothesis test of the mean, goodness of fit test, Bayesian credible interval and simple linear regression modal will be used to complete the report. The steps will be, suppose house price follows exponential distribution and calculate λ through MLE, then applying 90% confidence interval to mean prices. Moreover, the null hypothesis for mean price of old houses $H_0: \mu = 505000$, the alternative hypothesis for mean price of old houses $H_a: \mu \neq 505000$ and the null hypothesis for the mean price of new houses $H_0: \mu = 570000$, the alternative hypothesis for

mean price of new houses $H_a: \mu \neq 570000$. The report uses hypothesis test to test these assumptions. Notice that know the distribution of age of houses are also important. Hence, the report assumes $H_0: X \sim \text{Poi}(50)$ for the age distribution, then assuming we have prior distribution $Beta(55000{,}59000)$ for proportion of new houses and $Beta(420000{,}59000)$ for proportion of old houses and apply 90% Bayesian credible interval to the proportion of new houses and old houses.
Overall, we apply the methods to the clean dataset and then get results. Combining with the numerical summaries and graphs to analysis. Then, we approach the aim of the report.

Data

```r
# import data
HousePrice_eva <- read_csv("kc_house_data 2.csv")

# clean data
# move missing value by using filter() from Price, sqft_living and yr_built
# keep Price, sqft_living and yr_built
# find the number of old house and new house by using group_by()
HousePrice_clean <- HousePrice_eva %>%
  mutate(Old_New = ifelse(yr_built > 1971, "New", "Old"))%>%
  mutate(Age = 2021-yr_built) %>%
  mutate(Che_Expen = ifelse(price < 500000, "Cheap", "Expensive"))%>%
  filter(!is.na(price) & !is.na(yr_built) & !is.na(Old_New))%>%
  select(price, yr_built, Old_New, Age, Che_Expen)
Price_Old <- HousePrice_clean %>%
  filter(Old_New == 'Old')
Price_New <- HousePrice_clean %>%
  filter(Old_New == 'New')
```

The original dataset is called “HousePrice_eva”. There are 21613 observations and 21 variables in it. Notice that only part of them will be used. After cleaning, a new dataset which is named “HousePrice_clean” born. “HousePrice_clean” contains 21613 observations and 5 variables which are “price”, “yr_built”, “Che_Expen”, Age" and “Old_New”. Only these 5 variables have relationship with the topic.
I use filter() to remove missing values(learn code from R documentation, n.d.) and mutate to form new variables “Age”, “Old_New” and “Che_Expen.”Age" is formed by using 2021 minus the year a house built and I put the house that built before 1971 into “Old”, others are in “NEW”. That’s because there are already 50 years pass from 1971, these houses can be seen as old houses. Moreover, the price which is higher than 102594 dollar will be marked as expansive and others are cheap. That’s because the median King County Household Income was 102594 dollar in 2019(King County Economic Indicators, n.d. General Indicators). Notice that buying a house when just working for one year seems not realistic, but after five years, people should have deposit and maybe begin to consider to buy a house. Hence, the report assumes 500000 dollar is the standard for judging a house is cheap or expansive. Lastly, the group for old and new and the group for cheap and expansive are independent which means I decide houses into old and new without consider

if one house is cheap or not. The same condition applies for dividing houses into cheap and expensive. Here is the important definition of each variables(House Sales in King County, USA, 2016, Discussion):

Name	Description
price	Price of each home sold
Che_Expen	The price of houses are cheap or expensive
yr_built	The year the house was initially built
Old_New	The houses are New or Old
Age	The age of houses

Numerical Summaries

```
# Calculate some numerical data via R
summary(HousePrice_clean$price)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   75000  321950  450000  540088  645000 7700000

summary(HousePrice_clean$Age)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    6.00   24.00   46.00   49.99   70.00  121.00

summary(Price_Old$price)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   75000  293000  425000  505157  605000 7700000

summary(Price_New$price)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   90000  344938  478000  571803  675000 6885000

HousePrice_clean %>%
  summarise(prop_new = sum(Old_New=="New")/n())

## # A tibble: 1 x 1
##   prop_new
##      <dbl>
## 1    0.524

HousePrice_clean %>%
  summarise(prop_cheap = sum(Che_Expen=="Cheap")/n())

## # A tibble: 1 x 1
##   prop_cheap
##        <dbl>
## 1      0.574
```

```
HousePrice_clean %>%
  group_by(Old_New)%>%
  summarise(n=n())
```

```
## # A tibble: 2 x 2
##   Old_New     n
##   <chr>   <int>
## 1 New     11328
## 2 Old     10285
```

```
HousePrice_clean %>%
  group_by(Che_Expen)%>%
  summarise(n=n())
```

```
## # A tibble: 2 x 2
##   Che_Expen     n
##   <chr>     <int>
## 1 Cheap     12408
## 2 Expensive  9205
```

Here is the numerical summaries for “price”, “sqft_living” and “Age”:

Name	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
price	75000	321950	450000	540088	645000	7700000
———	———–	–	–	–	–	—
Age	6.00	24.00	46.00	49.99	70.00	121.00
———	———–	–	–	–	–	—
price for old	75000	293000	425000	505157	605000	7700000
———	———–	–	–	–	–	—
price for new	90000	344938	478000	571803	675000	6885000

The table shows that the min price is 75000 dollar and the highest price is 7700000 dollar. The mean is 540088 dollar and median is 450000 dollar. Also, the 1st quantile is 321950 dollar, which means there are 25% houses are cheaper than 321950 dollar. In addition, there are 25% houses are expensive than 645000 dollar. Lastly, the range is obvious which is 7700000-75000=7625000 dollar. This range is large.
Also, the min age is 6 years old and max is 121.o years old. The mean is 49.99 years old and the median is 46.00 years old. In addition, there are 25% houses are younger than 24.00 years old and 25% houses are older than 70.00 years old.
Moreover, The min and max price and range for old houses is the same as the min price for all houses. This means the price of old houses has larger range than the new. In addition, the mean for the old is 505157 dollar and the median for the old is 425000 dollar. The 1st quantile is 293000 dollar and there are 25% houses are expensive than 605000 dollar. All

of them are smaller than the new, which min is 571803 dollar, median is 478000 dollar 1st Quantile is 344938 dollar and 3rd Quantile is 675000 dollar.
By using group_by(), I got the proportion of new houses and using 1 to minus it, i got the proportion of new house. Also, I got the proportion of cheap houses and expensive houses.

Name	proportion
Old	0.475871
——	——–
New	0.524129
——	——–
Cheap	0.5740989
——	——–
Expensive	0.4259011

The proportion of old is 0.475871 and new is 0.524129. Also, The proportion of cheap houses is 0.5740989 and the old is 0.4259011.
Here is the number of old houses, new houses, expansive houses and cheap houses:

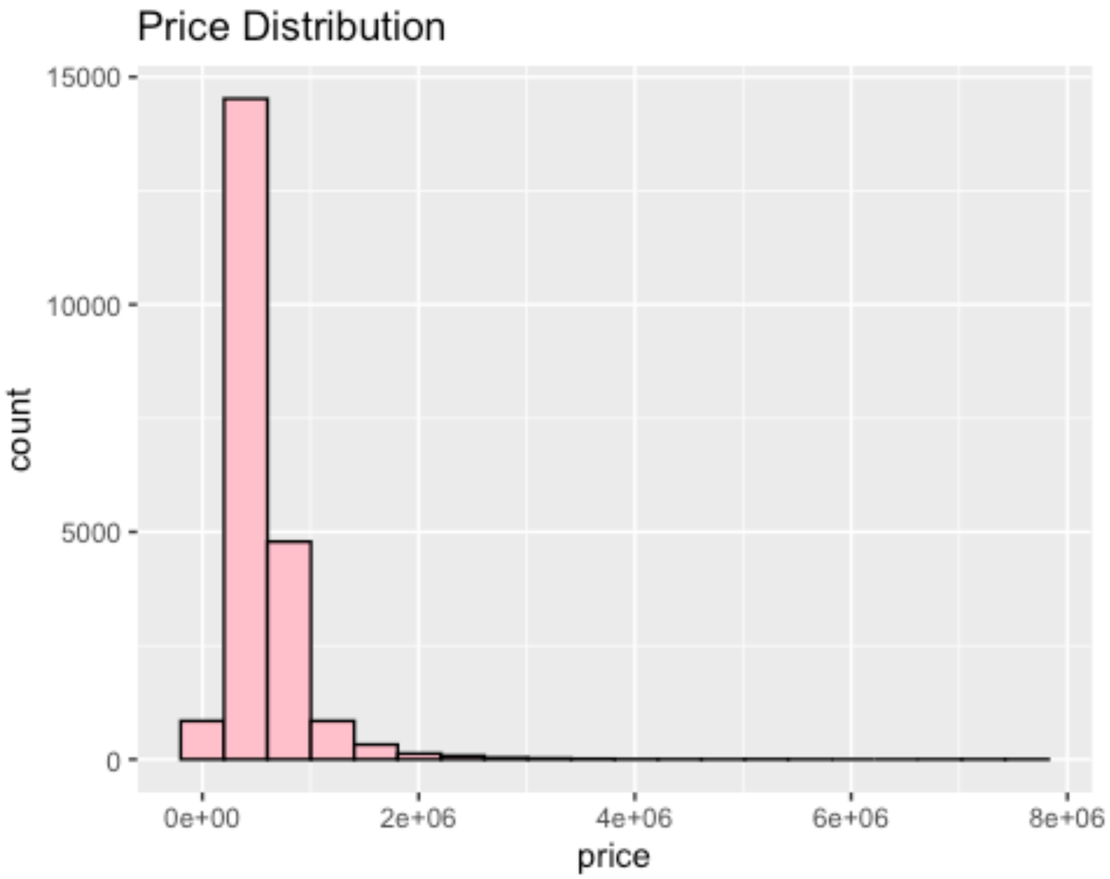
Name	Number
Old	10285
——	——–
New	11328
——	——–
Cheap	12408
——	——–
Expensive	9205

There are 10285 old houses, 11328 new houses, 12408 cheap houses and 9205 expensive houses. The differences between old houses and new houses is 11328-10285 = 1043. The differences between old houses and new houses is 12408-9205 = 3203. These differences are not that large.

Graphical Summaries

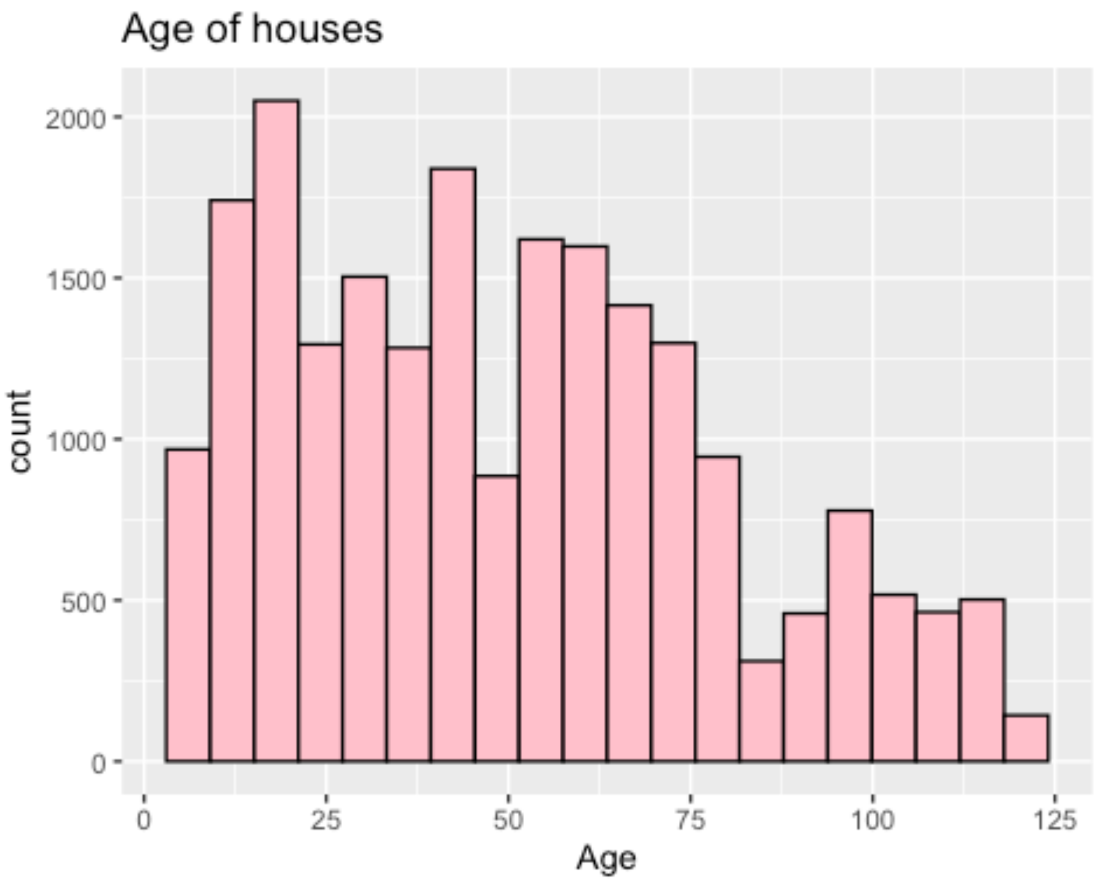
Boxplots, histograms and barplots shows the numerical value above clearly.

```
#Create side-by-side boxplot to for price, living spaces and age of houses
#Create a barplot to show the number of old houses and new houses
ggplot(data = HousePrice_clean, aes(x=price)) +
  geom_histogram(fill='pink',color='black', bins=20) +
  labs(x="price", title="Price Distribution")
```



In the histogram of “price”, we can see it has unimodal, right-skewed distribution. It centered at about 500000 dollar. The min of it is below 100000 dollar and the max is about 7000000 dollar.

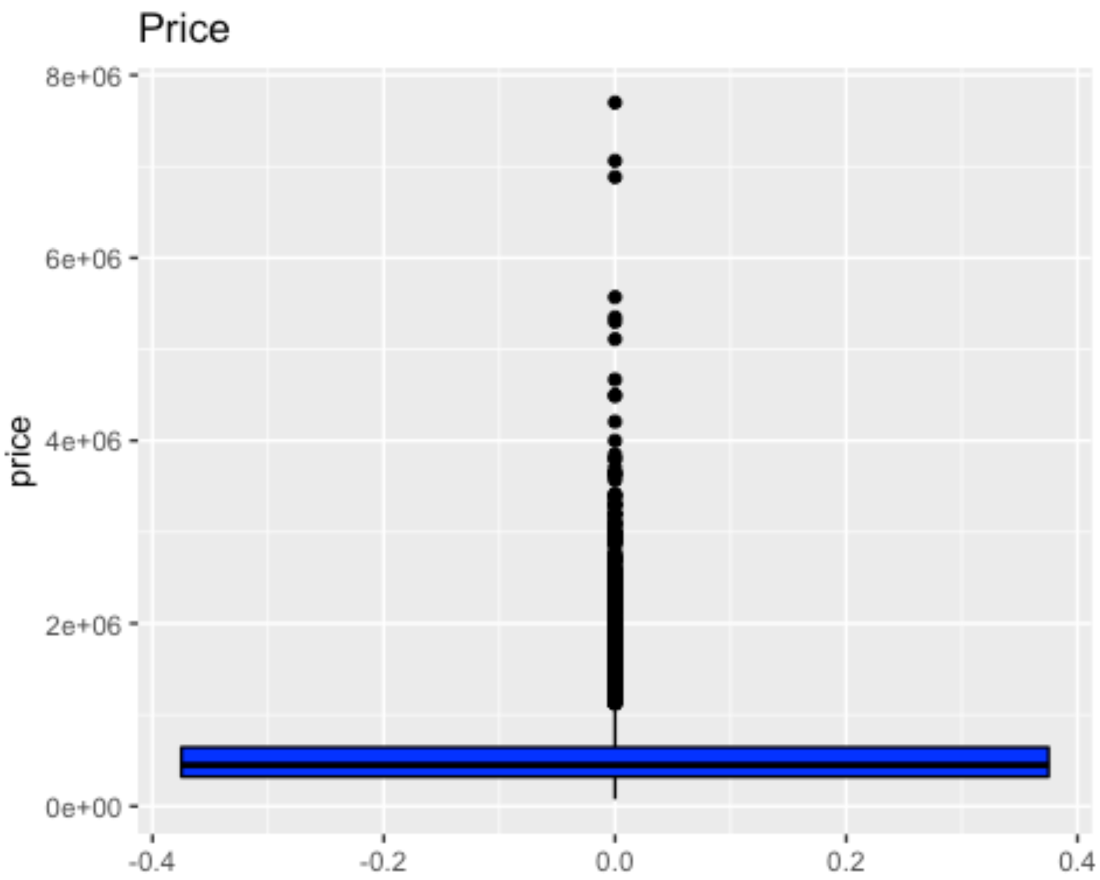
```
ggplot(data = HousePrice_clean, aes(x=Age)) +
  geom_histogram(fill='pink',color='black', bins=20) +
  labs(x="Age", title="Age of houses")
```



In the histogram of “Age of houses”, we can see it has biomodal, right-skewed distribution. It has modals at about 20 and 90. The min is about 5 years and the max is about 122 years.

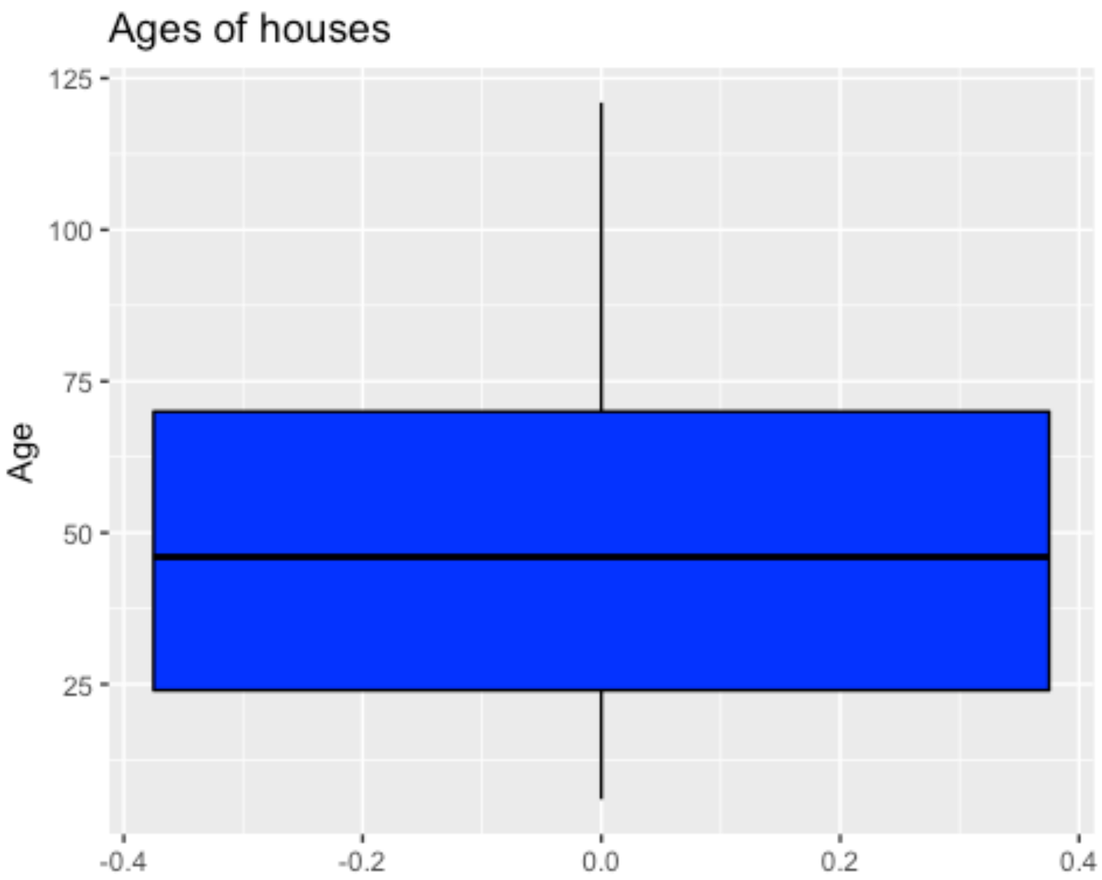
In order to see the results more clearly, I use boxplot to help.

```
ggplot(data = HousePrice_clean, aes(y=price)) +
  geom_boxplot(fill='blue', color='black') +
  labs(y="price", title="Price")
```



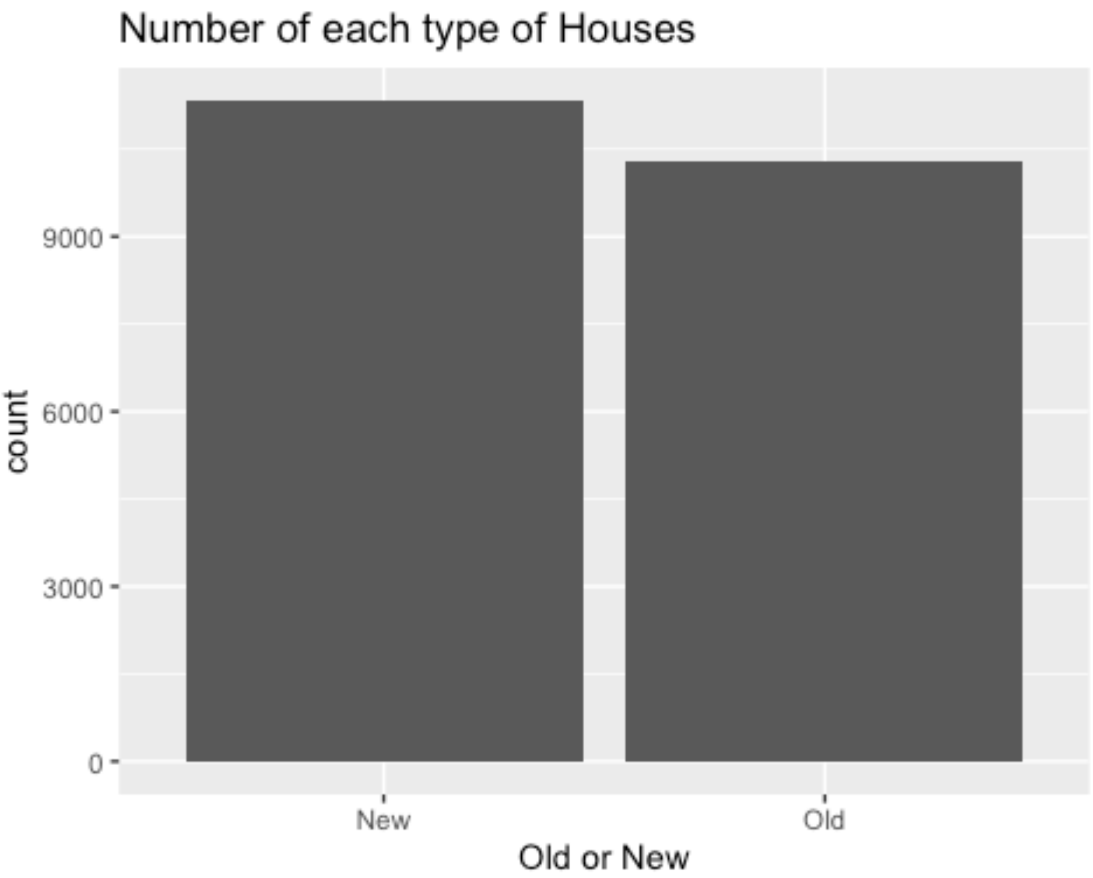
For price, it is a right-skewed distribution and centered at about 500000 dollar. The min is about 80000 dollar and the max is about 7500000 dollar. Hence, the range is about 7420000. Moreover, Q1 is about 300000 dollar and Q3 is about 600000 dollar. It has many outliers which are from 1000000 dollar to 7500000 dollar.

```
ggplot(data = HousePrice_clean, aes(y=Age)) +
  geom_boxplot(fill='blue', color='black') +
  labs(y="Age", title="Ages of houses")
```



For Age, it is a right-skewed distribution and median at 47 years old. The min is about 5 years old and the max is about 120 years old. Hence, the range is about 100 years old. Moreover, Q1 is about 21 years old and Q3 is about 70 years old.
Also, A barplot is needed to show the number of old houses and new houses.

```
ggplot(data=HousePrice_clean, aes(x=Old_New)) +
  geom_bar() +
  labs(x="Old or New", title="Number of each type of Houses")
```



In the barplot, one can see the total number of houses is 21613. About 11000 of them are new and 10000 of them are old. There are about 1000 differences between them.
All these data meets the numerical value above.

Methods

There are 6 methods will be used which are maximum likelihood estimator derivation(MLE), confidence interval via empirical bootstrap, hypothesis test of the mean, goodness of fit test, Bayesian credible interval and simple linear regression modal. The methods are ordered as above. Since the report are interesting on houses prices and ages. All these methods explore mean, distribution and etc. which have relationship with houses prices and ages. Recall the report has divided data into based on "yr_built" or based on "price". Hence, the report plans to use maximum likelihood estimator derivation and confidence interval via bootstrap to explore mean, distribution and etc. for data about price. Then, hypothesis test is used to check assumption of mean old houses price and mean new houses price. Lastly, goodness of fit test and Bayesian credible interval are used to explore distribution and etc. for data about age. Lastly, the report will give a relationship between Houses prices and ages by using simple linear regression modal.

Maximum Likelihood Estimator

Maximum Likelihood Estimator(MLE) is used to estimating an unknown parameters of a probability distribution. In order to estimate, we need to assume what kind of distribution the data follow by drawing a histogram, and then derive MLE(see appendix) so that we can calculate the parameters.
The report uses MLE to estimate the parameters of the probability distribution that houses price follow. The report assumes the data is a random sample of Exponential random variable with λ. Exponential distribution in used to consider the time between events. And we use it to consider the money that people spend on a house here. There are three reasons for estimating the data follows Exponential distribution. Firstly, only continuous data may follow Exponential distribution and price is continuous data. Secondly, the future probabilities does not influence by the past, and price is this kind of data. Thirdly, the histogram above shows that the data may follows Exponential distribution.
The report has used the MLE approach to estimate the λ. MLE derivations are in the Appendix.

Confidence Interval

The report uses 95% Confidence Interval(CI) to explore the mean price of houses through bootstrap sampling.The bootstrap sampling is a method that used to estimate the sampling distribution by re-sampling from the original sample. Notice that we sampling random and with replacement. I use this kind of method is because we donot have enough group of data to do re_sampling from the population. In other words, I use bootstrap to create bootstrap samples through the data we have so that we can complete the estimation. In addition, I choose to use CI here is because it is important to show how much confidence one have to be sure that the value is within this range and this is helpful for analyzing.
The report chooses to use empirical bootstrap sampling. That's because the empirical bootstrap makes no assumptions regarding the distribution of the sample. Although we have estimate the distribution the price follow in MLE subsection, there maybe some errors in estimation. Hence, using empirical bootstrap sampling will be more rigorous. Moreover, the report also rejects the option that is using Z/t approach. Although the sample size seems large, it is small compares the number of houses all over the world.
Treating the data we have as sample data and then randomly take the bootstrap samples with replacement from the sample data. Notice that the size of sample data and bootstrap data should be the same. Repeating the step for 2000 times and then calculating the statistic in every bootstrap sample, collecting and drawing the results into one histogram which is used to show the distribution of the bootstrap sampling.
Our goal is to explore the variability of estimates, so the report will not say the data set we get from bootstrap sampling is better than the original data set.
The report uses 90% confidence interval. The middle 90% of values of the bootstrap statistics form a 90% confidence interval for the parameter. The 90% confidence interval means we are 90% confident that the real parameter is in our interval. The parameter here will be the mean price of houses in King County and we choose to use 90% as the confidence level is because we have many sample so the confidence interval should be

narrower(Dr. Saul McLeod, June 10th, 2019, Paragraph 2) than usual.

Hypothesis Test for mean age of houses

The hypothesis test is used to test an assumption of sample data regarding a population parameter(CHRISTINA MAJASKI, oct. 24, 2020, paragraph 1).
The report apply this method on mean price of old houses and new houses. I choose to use this method is because testing the assumption for the mean prices is appreciate for using Hypothesis Test. Also the results are helpful for having insights on the relationship between prices and ages.The null hypothesis for mean prices of new house is $H_0: \mu = 505000$ and the alternative hypothesis is $H_a: \mu \neq 505000$. Also, the null hypothesis for mean prices of new house is $H_0: \mu = 570000$ and the alternative hypothesis is $H_a: \mu \neq 570000$.
Lastly, I decide to choose $\alpha = 0.05$ as the significance level, because it is the most common choice. Although there is no scientific judgement can be access for why $\alpha = 0.05$ is the most common choice, follow the majority is the most safe way.

Goodness of Fit Test

Goodness of fit test is used to test an assumption of what kind of distribution a sample data follows(Stephanie Glen, Nov. 7, 2014, paragraph 1).
The report chooses to use this method to test if the number of different age of houses follows Poi(50). Poisson distribution is used to show probability that events occur during a time. The reason for choosing $Poi(50)$ is because age is discrete and numerical variable. Also, since we separate houses into old and new by using 50 years old as standard, choosing $Poi(50)$ is reasonable. The reason for choosing this method is because testing distribution for the houses' age is appreciate for using Goodness of Fit Test .
Here are some important definitions.

Name	Description
X	number of houses houses

We use "X" to represents number of houses.
The null hypothesis is $H_0: X \sim$ Poi(50). and the alternative hypothesis is $H_a: X$ does not follow Poi(50).
The significance level is the same as the one for hypothesis test.

Bayesian Credible Interval

Bayesian Credible Interval is used to describe the the unknown parameters' uncertainty which one people tries to estimate(Dominique Makowski, Daniel Lüdecke, Mattan S. Ben-Shachar, Indrajeet Patil, Michael D. Wilson, 2018, Paragraph 1). The report uses this method is because we donot know what parameter the distribution follows but this is

important for analyzing. Hence, applying Bayesian Credible interval is appreciate.
Here are some important definitions:

Name	Description
n	total number of houses
X1	number of new houses
p1	Proportion of new houses
X2	Proportion of old houses
p2	Proportion of old houses

We use "n" to represents the total number of houses, "X1" to represents number of new houses, "X2" to represents number of old houses, p1 to represents the proportion of new houses and "p2" is proportion of old house. So there are two Bayesian Credible Interval will be demonstrated.
The report are interested in finding a 90% credible interval of the parameter p. I assume the data is a random sample which follows $X \sim \text{Bin}(n, p)$; and the prior distribution for p1 is $p \sim \text{Beta}(55000{,}59000)$ in hopes of yielding a non-informative prior and he prior distribution for p1 is $p \sim \text{Beta}(42000{,}59000)$. Notice that the number in prior distribution is random choices.
Thus, a range of value is derived by using the 5th and 95th percentiles of this distribution. This means p has 90% probability of falling within this range. The report choose to use 90% as the confidence level is because the confidence interval should be narrower when we have such many data(Dr. Saul McLeod, June 10th, 2019, Paragraph 2) than usual.
All the derivations can be find in Appendix.

Linear Regression

The report assumes there is a dataset with X as non-random independent variable and the y are realizations of random variables Y satisfy the following simple linear regression model, notice that U_i are independent random variables with $\text{E}[U_i] = 0$ and $\text{Var}(U_i) = \sigma^2$.(Week3-LinearRegressionModels, page 6).
Thus, our true model is

$$Y_i = \alpha + \beta x_i + U_i$$

Name	Result
Y_i	dependent variable
——————	——–
$\alpha + \beta x_i$	population regression line
——————	——–
α	population Y intercept

β	slope
X_i	independent variable
U_i	error

```
#check if it is numeric
check_age <- HousePrice_clean
class(check_age $ Age)

## [1] "numeric"

check_prices <- HousePrice_clean
class(check_prices $ price)

## [1] "numeric"
```

Function class()(Learn from Factor in R: Categorical Variable & Continuous Variables, n.d.)is used to check that if price and Age are numerical variable. The results are “numeric”. Hence, using simple linear regression modal is appreciate.

Results

The report uses methods to explore relevant results for our topic. Overall, we find the price of houses follows price of houses$\sim$ Exp$(1.85155e - 06)$ through MLE. Then, by using bootstrap to formulate 90% confidence interval, we say we have 90% confident that the mean of houses price is between 534967.1 dollar and 544891.9 dollar. After having an overall insights of prices, we move to age of houses. The p values which get through Hypothesis test shows that we have no evidence to against H_0. Also, by using credible interval, we are 90% confident that the proportion of new houses is between 0.5229628 and 0.5272637. In addition, finding the distribution for proportion of old houses are also important. By using Goodness fit test, …Lastly, we use linear regression to give a view of the relationship between houses price and age. That’s there is a weak linear relationship and the data is actually not fit the linear regression modal.

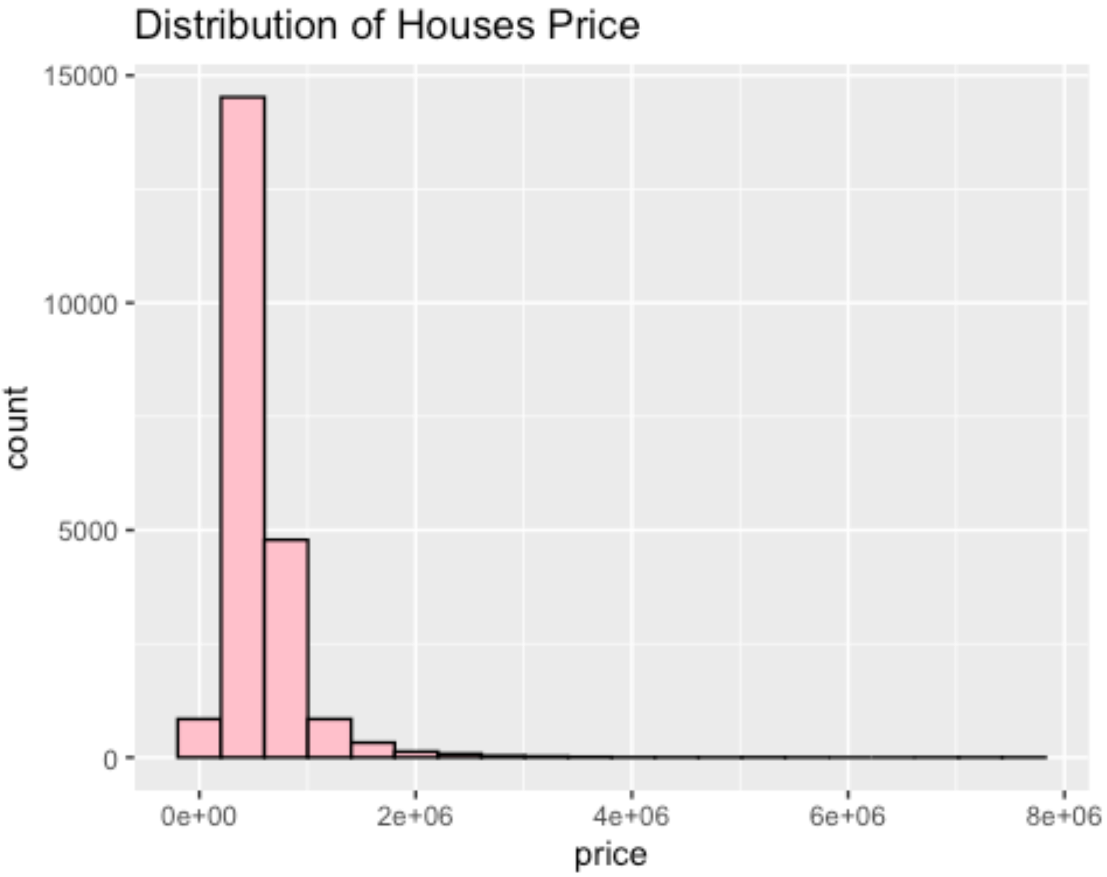
Section 1: Exploring the relevant information for price of houses

Maximum Likelihood Estimator

The report use MLE to estimate the distribution of houses’ price.

```
#Draw a histogram for price
# finding what kind of distribution the price of houses follow
```

```
HousePrice_clean %>%
  ggplot(aes(x=price)) +
  geom_histogram(fill='pink',color='black', bins=20) +
  labs(x="price", title="Distribution of Houses Price")
```



```
set.seed(368)
lambda_hat <- 1/mean(HousePrice_clean$price)
lambda_hat

## [1] 1.85155e-06
```

By drawing the histogram, the report assumes the distribution of houses price is houses price $\sim$ Exp(λ).
By deriving(see Appendix), we get the function for MLE of Exponential Distribution which is $\hat{\lambda} = \frac{1}{\bar{x}}$(Notice that $\bar{x}$ is sample mean).
The result shows that the distribution of houses price follows houses price $\sim$ Exp($1.85155e - 06$).

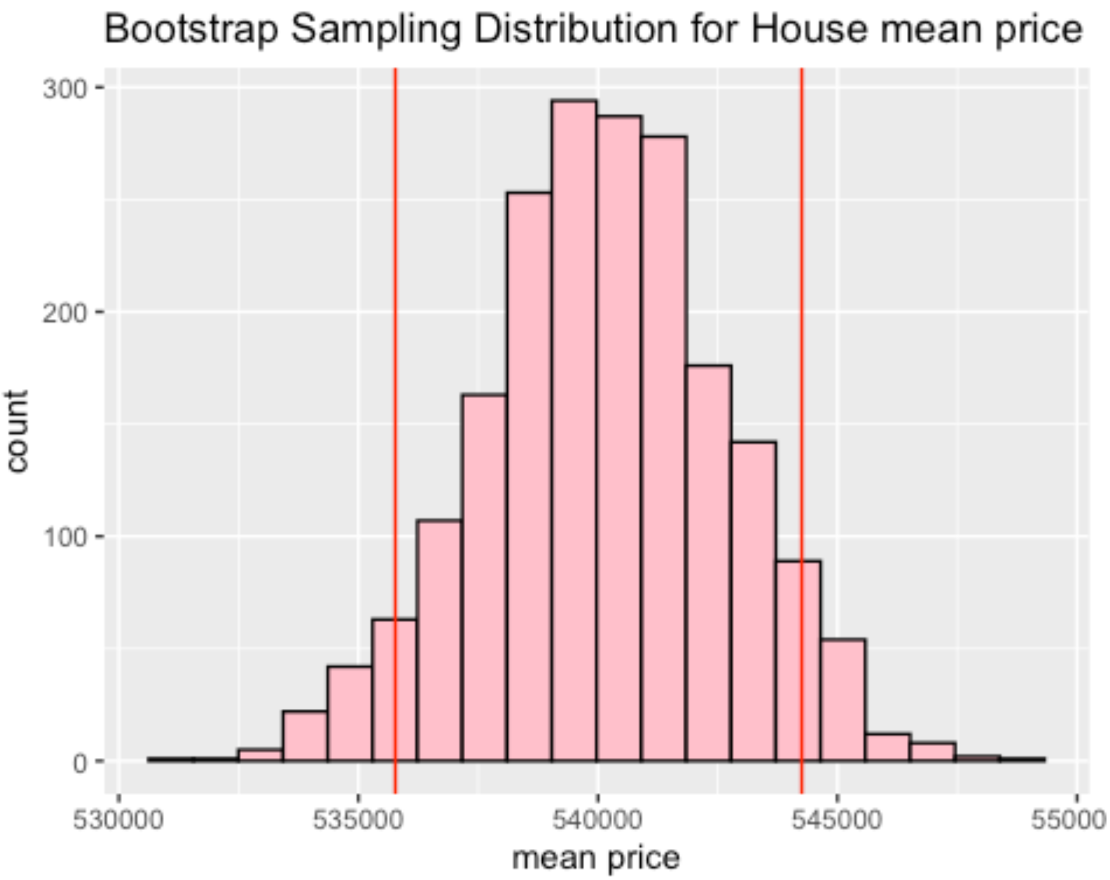
Confidence Interval

The report applies 90% Confidence interval on mean house price.

```r
set.seed(368)
# the number of simulations is 2000
repetitions <- 2000
# Create an empty vector to store the simulation results
HPprice <- rep(NA, repetitions)
n_obs <- nrow(HousePrice_clean)
for(i in 1:repetitions){
HP_samp <- sample_n(HousePrice_clean, size = n_obs, replace = TRUE)
HP_mean <- HP_samp %>%
summarise(mean_HP = mean(price)) %>%
as.numeric()
HPprice[i] <- HP_mean
}
HPprice <- tibble(mean_HP  = HPprice)
#Calulate the 5% quantile and 95% quantile.
quantile(HPprice$mean_HP, c(0.05,0.95))
```

```
##       5%      95%
## 535771.0 544253.9
```

```r
# Draw v line on the histogram.
HPprice %>%
  ggplot(aes(x=mean_HP)) +
  geom_histogram(fill='pink',color='black', bins=20) +
  labs(x="mean price", title="Bootstrap Sampling Distribution for House mean price")+ geom_vline(xintercept = quantile(HPprice$mean_HP, c(0.05,0.95)), color = "red")
```



5% Quantile	95% Quantile
535771.0	544253.9

One can state that the 90% confidence interval for the mean price of houses is [535771.0,544253.9], which means we are 90% confident that the true population mean price of houses is between 534967.1 dollar and 544891.9 dollar. This corresponds to the numerical summary above, where the mean is about 540088 dollar. The confidence interval method fits the data.

Hypothesis Test

The report uses hypothesis test to test if we have evidence to reject the assumptions for old houses price and new houses price.

```
#Calculate p-values
x_bar <- mean(Price_Old$price)
x_bar

## [1] 505156.9
```

```
n <- nrow(Price_Old)
s <- sqrt(var(Price_Old$price))
test_stat <- abs((x_bar- 505000)/(s/sqrt(n)))
p_value <- 2*(1-pt(test_stat, df=n-1))
p_value

## [1] 0.96335
```

Name	Number
$\bar{x}$	505156.9
p-value	0.96335

The old houses mean price is 505156.9 years old. We get P-value as 0.96335. Since P-value is bigger than 0.05, we have no evidence to reject H_0. That's means we have no evidence to reject that the mean price for old houses is 505156.9 dollar.

```
#Calculate p-values
x_bar <- mean(Price_New$price)
x_bar

## [1] 571803.1

n <- nrow(Price_New)
s <- sqrt(var(Price_New$price))
test_stat <- abs((x_bar- 575000)/(s/sqrt(n)))
p_value <- 2*(1-pt(test_stat, df=n-1))
p_value

## [1] 0.3734346
```

Name	Number
$\bar{x}$	571803.1
p-value	0.3734346

The new house mean price is 571803.1 dollar for new houses and P-value is 0.3734346. Since P-value is bigger than 0.05, we have not evidence to reject H_0. That's means we have no evidence to reject that the mean price for new houses is 571803.1 dollar. Compared these two results, one can see there is about 70000 dollar difference. This difference isn't that significant.

Section 2: Exploring the relevant information for age of houses

Goodness of Fit Test

The report uses goodness of fit test to check if the assumption we made for the distribution of houses age is hold.

```
set.seed(368)
sum_x <- sum(HousePrice_clean$Age)
n <- nrow(HousePrice_clean)
lambda_hat <- sum_x/n
lambda_hat

## [1] 49.99486

likelihood_ratio <- exp(-n*(50-lambda_hat))*(50/lambda_hat)^sum_x
p_value <- 1-pchisq(-2*log(likelihood_ratio), df=1)
p_value

## [1] 0.9149639
```

Name	Value
lambda_hat	49.99486
———————	——-
p_value	0.9149639

I get the $\hat{\lambda}$, which is the mean age as 49.99486. Then I calculate the p value, which is 0.9149639. Since the p value is bigger than 0.05, we have no evidence to reject our H_0, which means we have no evidence to reject that X ~ Poi(50).

Bayesian Credible Interval

The report uses 90% Bayesian Credible Interval for proportion of new houses.

```
# Find the 5% lower credibal interval_lower and 95% upper credible interval.
set.seed(368)
n <- nrow(HousePrice_clean)
x <- sum(HousePrice_clean$Old_New=="New")
post_alpha <- n+55000
post_beta <- n-x+59000
cred_interval_lower <- qbeta(0.05, post_alpha, post_beta)
cred_interval_upper <- qbeta(0.95, post_alpha, post_beta)
cred_interval_lower

## [1] 0.5229628

cred_interval_upper
```

```
## [1] 0.5272637
```

lower credible interval	upper credible interval
0.5229628	0.5272637

The report assumes the random sample follows $X \sim \text{Bin}(n, p)$ with the prior distribution for p is $p \sim \text{Beta}(55000{,}59000)$. By calculation via R, we get 0.5229628 for lower credible interval and 0.5272637 for upper credible interval. Hence, the proportion of new houses has 90% probability to fall into this range. Recall the proportion of new houses in numerical summaries which is 0.524129. 0.524129 lies in this range. Bayesian Credible Interval is suitable.

```
# Find the 5% lower credibal interval_lower and 95% upper credible interval.
set.seed(368)
n <- nrow(HousePrice_clean)
x <- sum(HousePrice_clean$Old_New=="Old")
post_alpha <- n+42000
post_beta <- n-x+59000
cred_interval_lower <- qbeta(0.05, post_alpha, post_beta)
cred_interval_upper <- qbeta(0.95, post_alpha, post_beta)
cred_interval_lower

## [1] 0.4726888

cred_interval_upper

## [1] 0.4771776
```

lower credible interval	upper credible interval
0.4726888	0.4771776

The report assumes the random sample follows $X \sim \text{Bin}(n, p)$ with the prior distribution for p is $p \sim \text{Beta}(42000{,}59000)$. By calculation via R, we get 0.4726888 for lower credible interval and 0.4771776 for upper credible interval. Hence, the proportion of old houses has 90% probability to fall into this range. Recall the proportion of new houses in numerical summaries which is 0.475871. 0.475871 lies in this range. Bayesian Credible Interval is suitable.

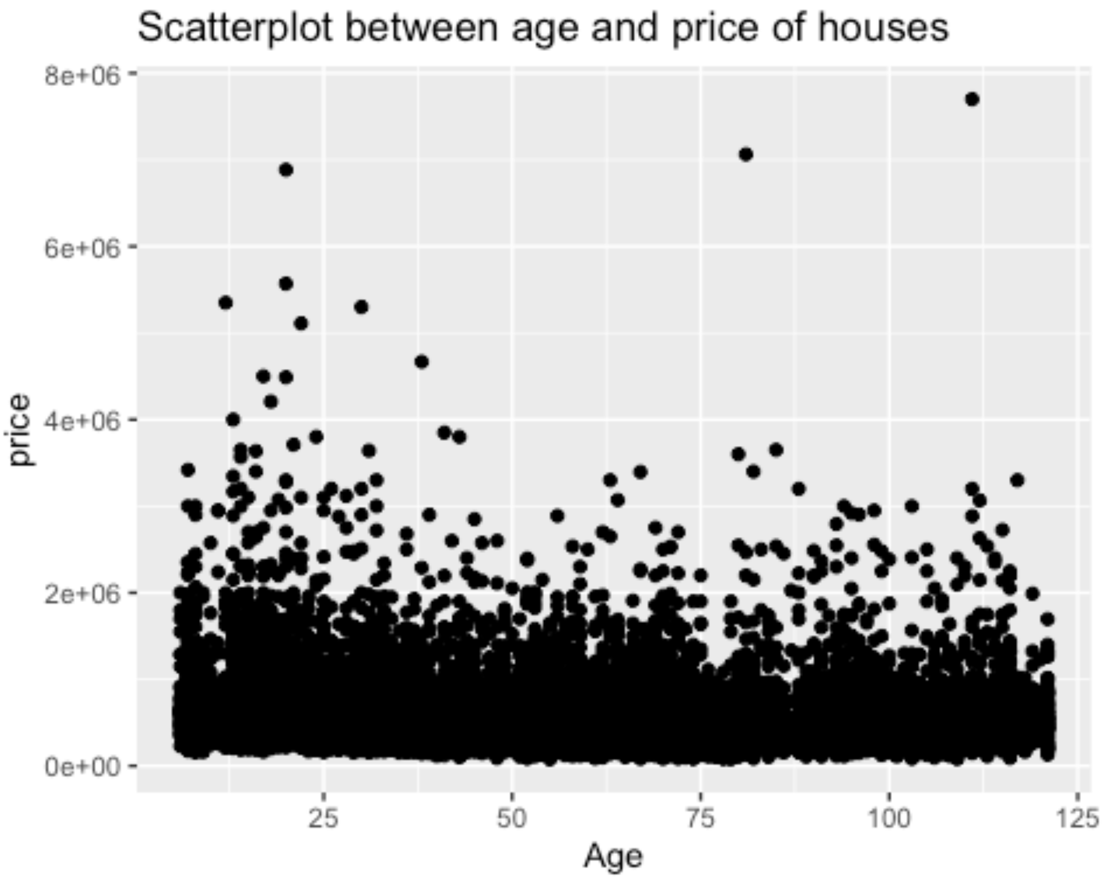
Section 3: Exploring the relationship between prices and ages

Linear Regression

The report uses simple linear regression to show the relationship between prices and ages.

```
# Create a Scatterplot between living spaces and prices of the house.
```

```
HousePrice_clean %>%
  ggplot(aes(x = Age, y = price)) + geom_point() + labs(x = "Age", y =
"price", title = "Scatterplot between age and price of houses")
```



```
Sp_Prices_model <- lm(price ~ Age, data = HousePrice_clean)
Sp_Prices_model

##
## Call:
## lm(formula = price ~ Age, data = HousePrice_clean)
##
## Coefficients:
## (Intercept)          Age
##    573838.2       -675.1
```

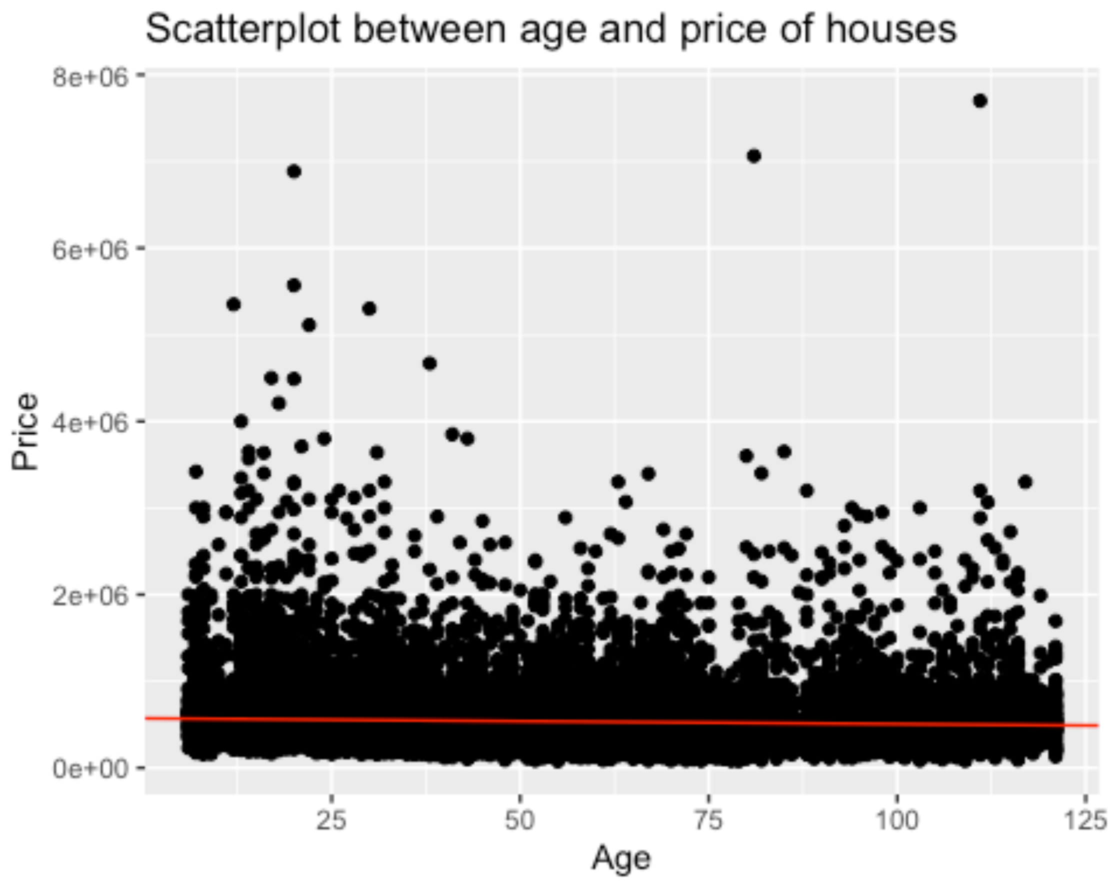
There is a negative weak linear relationship between age and price of the houses. When the age of houses become larger, the price become lower. There are some outliers, which there is two houses that has ages over 75 years but sold for price higher than 7000000. This price are even higher than the new(age is less than 50 years).

Name	Result
$\hat{\alpha}$	573838.2

$\hat{\beta}$	-675.1
X_i	Age
$\widehat{Y_i}$	573838.2 + (-675.1)X_i
U_i	error

The table shows that the the estimated model is $\widehat{Y_i}$ = $\hat{\alpha}$ + $\hat{\beta}X_i$, which is 573838.2 + (-675.1)X_i. The independent variable is ages and the dependent variable is expected houses price, which calculates by formula 573838.2 + (-675.1)X_i.
The y intercept of the estimated regression line is 573838.2, which means a 0 years houses can be sold with 57838.2 dollar. It actually make sense to a large extent, because the max price of houses are 7700000 dollar and mean price are 540088 dollar. According to the linear regression, the new usually have higher price than the old. Hence, a 0 years new house has a price which is higher than the mean price is reasonable. However, the expected y intercept will be more reasonable if the value can be higher since 0-year-old houses are really new. This situations happen may because there are other factors are influencing price(see more in limitations).
Moreover, the slope of the estimated regression line is -675.1, which means when the building age increase by one year, the price will decrease 675.1 dollar. It is make sense because as a building become old, there will be some issues happen in the buildings and the buildings may have a bad appearance. These makes the houses have less competitiveness with the new houses. Hence, the price will be lower. However, since the age only increase one year, the house will not depreciate too much. Thus, -675.1 dollar is reasonable.

```r
# Adding the regression line to the scatterplot
HousePrice_clean %>%
  ggplot(aes(x = Age, y = price)) + geom_point() + labs(x = "Age", y =
"Price", title = "Scatterplot between age and price of houses") +
  geom_abline(slope = -675.1, intercept = 573838.2, col ="red")
```



The red line is the estimated regression line. Hence, points on this line are follow the estimated model. There will always many factors influence the price since our data is random and realistic. These factors make points located beside the line. And these points can be counted as errors.

From the graph, we can see the red line is almost a horizontal line since the $\hat{\beta}$ are small. Also, points are not evenly located beside the line. There are more points above the line. Hence, there is an increasing variance and forms a fan pattern. All these make the data not fit the $\mathrm{E}[U_i] = 0$ and $\mathrm{Var}(U_i) = \sigma^2$. Thus, these data are not suitable for simple linear regression modal.

Conclusions

The aim of the report is to analyzing the relationship between house prices and ages in King County, USA, and give advice to people who want to sold their houses.

"New houses have higher price and selling houses when it is new will give more profit to the sellers." is the main hypothesis for this report. All methods and analysis will be made around it. In order to check our main assumption, many other hypothesis based on data have been made. Firstly, the report assumes the price of houses $X \sim \mathrm{Exp}(\lambda)$. Also, we apply assumption for old houses mean price and new houses mean price, too. Moreover, we assume H_0 is the ages of houses $X \sim \mathrm{Poi}(50)$ and H_a: X does not follow $\mathrm{Poi}(50)$. For old houses mean price, $H_0 = 505156.9$ and $H_a \neq 505156.9$ and for new houses mean price, $H_0 = 571803.1$ and $H_a \neq 571803.1$. Lastly, we assumes the proportion of old houses and new houses are both follows binomial distribution. Be more specific, we assume the proportion of new houses' prior distribution $p \sim \text{Beta}(55000,59000)$ and the proportion of old houses' prior distribution $p \sim \text{Beta}(42000,59000)$.

In order to make an analysis, we uses methods to calculate results, test our hypothesis and show relationships. In order to reduce confusion, I separate the results into three parts. In section 1, we show that the price of houses $X \sim \text{Exp}(1.85155e - 06)$ through maximum likelihood estimator, and we have 90% confident that the mean price of houses is between 534967.1 dollar and 544891.9 dollar by using empirical bootstrap sampling. Also, through using hypothesis test, we show that there is no evidence to against the H_0 for both old houses mean price and new houses mean price since the p values are all smaller than 0.05. That's means we have evidence to support that the mean price for new is 571803.1 dollar and the price for old is 505156.9 dollar. In section 2, we firstly show that there is no evidence to against our H_0 since the p values are also smaller than 0.05. Hence, we have evidence to support that the mean age is 49.99486 years old. When it comes to the binomial distributions that are followed by the proportion of old houses and new houses, we say there are 90% probabilities for the proportion of new houses to lie between 0.5229628 to 0.5272637 and 90% probabilities for the proportion of old houses to lie between 0.4726888 and 0.4771776 by finding 90% Bayesian credible interval. Compared the interval for the proportion of new and old, we can find there is no huge difference between the proportion. In Section 3, we use the simple linear regression modal to show there is a weak relationship between price and age in King county and further analyse that the data is not suitable for linear regression modal since it does not satisfy $\text{E}[U_i] = 0$ and $\text{Var}(U_i) = \sigma^2$. These are key results we have in the report. They are key evidence for helping sellers predict future market conditions.

Through understanding all the steps and results above, one can understand the conditions of houses price and age in King County. One can uses the Exponential distribution of houses price to estimate the probability of different prices occur in the market and uses the Poisson distribution to estimate the number of houses in a certain age in the market. Combined these two distributions with the mean house age in King County, they will have a general idea for price they should give to their houses if they want to sold quickly. Be more specific, most of houses in King County has price about 500000 dollar, so selling in a price under this amount is a good choice. However, that's not enough. The report improves the accuracy through providing more details for old houses and new houses. By testing the hypothesis, we show the mean price for old houses and new houses. Compared with the proportion of old houses and new houses which are showed by Bayesian credible interval, one can plan to give a price about 571803.1 dollar if their houses is new or about 505156.9 dollar for their old houses since the proportion for old and new is almost equivalent, no such a situation that housing prices can be set high due to too few houses in the market can occur. Lastly, although the simple linear regression modal not fit for the data, one can still see some weak relationship between ages and houses. This implies that the value of the houses are still depreciating. So selling houses when they are new as much as possible.

The analysis above show the conditions in King County, when we move to the global market, there are differences and similarities. But the report can be used as a consultation.

Weaknesses

Firstly, we assume the mean price of houses follows Exponential distribution. However, that's an assumption according to the histogram, we lack other information to support our assumption. The reality is the mean price may not follow this distribution, so our assumption is not correct. Hence, the following calculation and analysis may not be able to give valuable advice. The same condition apply to all the assumptions the report have made.
Secondly, the data is limited in king County, USA and the data shown in the dataset are all surveyed between 2014-2015. Hence, we have limitations in time and location. Hence, the result we get may not suitable for globally.
Thirdly, the data is actually not suitable for simple linear regression modal. Hence, we cannot give a satisfied answer on the type of relationship between age and price. We may need a more sophisticated statistical methods to get the answer. This situation occurs becuase there are many other factors can influence the relationship such as location of houses.

Next Steps

For next step, we tend to find data in other locations and analysis the conditions in different areas. by comparing them, one can get a more scientific result and a deeper view on houses age and prices. So that, they can make more accurate decisions.
Also, finding more evidence to support hypothesis the report make will make the report seem more reasonable.
Lastly, the reference value of new data is greater, so try to find new data in the future.

Discussion

Although there are some limitations in this report, by using six methods and different graphs, we interpertate the house market in king County and give the auidance some valuable advice on how to sold their houses out. All analysis for this report was programmed using `R version 4.0.4`.

Bibliography

1. Grolemund, G. (2014, July 16) *Introduction to R Markdown*. RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: January 15, 2021)

2. Dekking, F. M., et al. (2005) *A Modern Introduction to Probability and Statistics: Understanding why and how.* Springer Science & Business Media.

3. Allaire, J.J., et. el. *References: Introduction to R Markdown*. RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: January 15, 2021)

4. King County Economic Indicators - King County. (2021). https://www.kingcounty.gov/independent/forecasting/King%20County%20Economic%20Indicators.aspx(Last Accessed: 17 April, 2021)

5. How Hypothesis Testing Works. (2021). https://www.investopedia.com/terms/h/hypothesistesting.asp(Last Accessed: 17 April, 2021) https://www.simplypsychology.org/confidence-interval.html

6. Goodness of Fit Test: What is it? - Statistics How To. (2021). https://www.statisticshowto.com/goodness-of-fit-test/(Last Accessed: 17 April, 2021)

7. Wickham, H. (n.d.). Filter. from https://www.rdocumentation.org/packages/dplyr/versions/0.7.8/topics/filter(Last Accessed: 17 April, 2021)

8. Mcleod, S. (2021). What are confidence intervals? | Simply Psychology. https://www.simplypsychology.org/confidence-interval.html(Last Accessed: 17 April, 2021)

9. House Sales in King County, USA. (2021). https://www.kaggle.com/harlfoxem/housesalesprediction(Last Accessed: 17 April, 2021)

10. Week3-LinearRegressionModels.pdf. (n.d.). https://q.utoronto.ca/courses/204754/files/11961833?wrap=1(Last Accessed: 17 April, 2021)

Appendix

Section 1

This section is for MLE. The report assumes the distribution of houses price is houses price $\sim$ Exp(λ).

The probability density function(pdf) is $f(x) = \lambda e^{-\lambda x}$. Step 1: find the Likelihood Function

$L(\lambda) \quad = f(x_1)\cdots f(x_n)$

$= \lambda e^{-\lambda x_1} \dots \lambda e^{-\lambda x_n}$ Step 2: Taking ln on $L(x) = f(x_1)\cdots f(x_n)$ to find $l(\lambda)$.

$= \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}$

$l(\lambda) \quad = \ln L(\lambda)$

$= \ln\left(\lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}\right)$ Step 3: Find first derivative, notice that we are estimating, so we

$= n\ln\lambda - \lambda \sum_{i=1}^{n} x_i$

use $\hat{\lambda}$ instead of λ:

$$\frac{dl}{d\lambda} = \frac{n}{\lambda} \quad - \sum_{i=1}^{n} x_i = 0$$

$$\frac{n}{\lambda} \quad = \sum_{i=1}^{n} x_i$$

$$\frac{1}{\hat{\lambda}} \quad = \frac{\sum_{i=1}^{n} x_i}{n}$$

Since $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

$$\hat{\lambda} = \frac{1}{\bar{x}}$$

Step 4: Do a second derivative test, check $\frac{d^2 l}{d\lambda^2}$ is bigger than 0 or not. Notice that, we have $\lambda = \hat{\lambda} = \frac{1}{\bar{x}}$, hence, the second derivative is respect to $\bar{x}$.

$-\frac{n}{\left(\frac{1}{\bar{x}}\right)^2} = -n\bar{x}^2.$

Since $\bar{x}^2$ is bigger than 0, $-n\bar{x}^2$ is smaller than 0. Then it is concave down. Our result is the maximum.

Thus the MLE is $\hat{\lambda} = \frac{1}{\bar{x}}$.

Also, the report test if the age of houses $\sim$ Poi(50). This need math deviation, too. The pdf for Poisson distribution is $P(x) = \frac{e^{-\lambda}\lambda^x}{x!}$.

Firstly, we calculate the the Likelihood Function:

$L(x) \quad = p(x_1)\cdots p(x_n)$

$$= \frac{e^{-\lambda}\lambda^{x_1}}{x_1!} \dots \frac{e^{-\lambda}\lambda^{x_n}}{x_n!}$$

$$= \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} x_i}}{x_1! \cdot \cdots \cdot x_n!}$$

Secondly, Taking ln.

$l(\lambda) = = -n\lambda + (\sum_{i=1}^{n} x_i)\ln\lambda - \ln(x_1! \dots x_n!)$

Thirdly, find first derivative with respective to λ:

$$\frac{\sum_{i=1}^{n} x_i}{\lambda} - n = 0$$

$$\frac{\sum_{i=1}^{n} x_i}{\lambda} = n$$

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Since $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

we get $\hat{\lambda} = \bar{x}$.

Fourthly, taking second derivative test:

$$\frac{d^2 l}{dx^2} = -\frac{\sum_{i=1}^{n} x_i}{\lambda^2}$$

$=-\frac{\sum_{i=1}^{n} x_i}{\bar{x}^2}$

Then the result is $= -\frac{n}{\bar{x}}$. Recall that n is the number of hoses and $\bar{x}$ is mean, so both of them are bigger than 0. Hence, it is max since it is concave down.

Then use this to get the Likelihood ratio(LR). The numerator is $L(50)$and $L(\hat{\lambda})$ is denominator

By $\frac{e^a}{e^b} = e^{a-b} = \frac{e^{-50n}50^{\sum_{i=1}^{n} x_i}}{x_1!\cdots x_n!}\frac{x_1!\cdots x_n!}{e^{-\bar{x}n}\bar{x}^{\sum_{i=1}^{n} x_i}}$

$= e^{-n(50-\bar{x})}\left(\frac{50}{\bar{x}}\right)^{\sum_{i=1}^{n} x_i}$

Then our p-value is $P(\chi_1^2 > -2\ln(LR))$.

Section 2

This section is for the beta posterior distribution.

Step 1: get the likelihood for binomial distribution: The likelihood for binomial is its pmf:$P(x) = \binom{n}{x} p^x (1-p)^{n-x}$, notice that, p is the proportion.

The pdf for beta is $\frac{\Gamma_(\alpha+\beta)}{\Gamma_{10}\Gamma_(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$ By formula, posterior is equal to L(p)f(p) which is $p^{x+\alpha-1}(1-p)^{n-x+\beta-1}$ after pluging in pmf and pdf. Hence, the posterior distribution is $BETA(X+\alpha\ n-x+\beta)$.