FINAL PROJECT

Alizeh Azhar

11/28/2021

Dataset Selection

The dataset that I have chosen for my final project is the *Housing dataset*, which contains *Sales Prices of Houses in the City of Windsor*. The dataset contains 546 number of observations and 12 variables. Out of the 12 variables, 6 are numeric and 6 are categorical. Table 1 below shows the summary of the dataframe.

Table 1: Data summary

Name	Housing
Number of rows	546
Number of columns	12
Column type frequency:	
character	6
numeric	6
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
driveway	0	1	2	3	0	2	0
recroom	0	1	2	3	0	2	0
fullbase	0	1	2	3	0	2	0
gashw	0	1	2	3	0	2	0
airco	0	1	2	3	0	2	0
prefarea	0	1	2	3	0	2	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
price	0	1	68121.60	26702.67	25000	49125	62000	82000	190000
lotsize	0	1	5150.27	2168.16	1650	3600	4600	6360	16200
bedrooms	0	1	2.97	0.74	1	2	3	3	6
bathrms	0	1	1.29	0.50	1	1	1	2	4
stories	0	1	1.81	0.87	1	1	2	2	4
garagepl	0	1	0.69	0.86	0	0	0	1	3

Exploratory Analysis and Data Summary

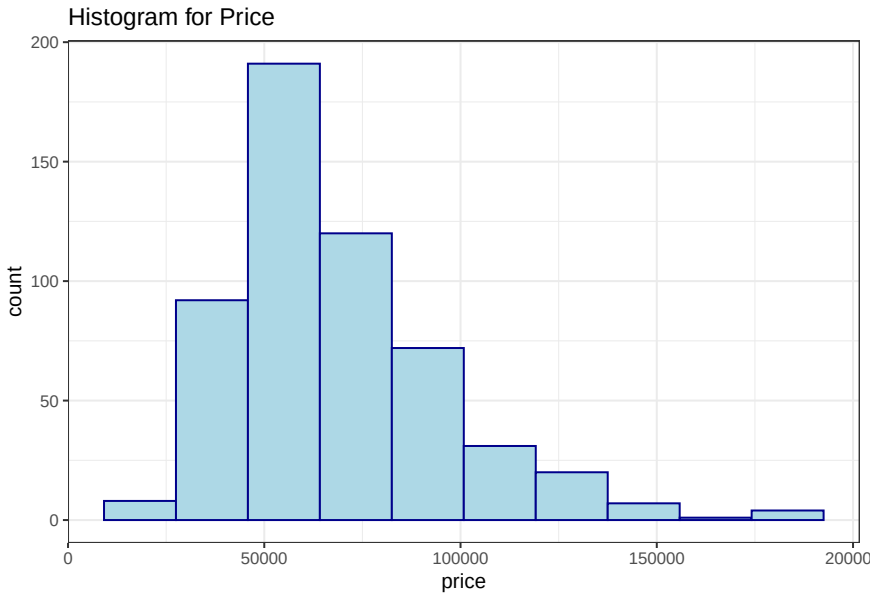
In this section, all the 12 variables in the dataset are summarized.

Price

It is a numeric continuous variable that represents the sale price of a house. In our analysis, it will be the dependent variable. The table below shows the summary statistics of the variable.

	Mean	Std.Dev	Min	Q1	Median	Q3	Max
price	68121.60	26702.67	25000.00	49000.00	62000.00	82000.00	190000.00

We can summarize the distribution of the `price` variable using a histogram.



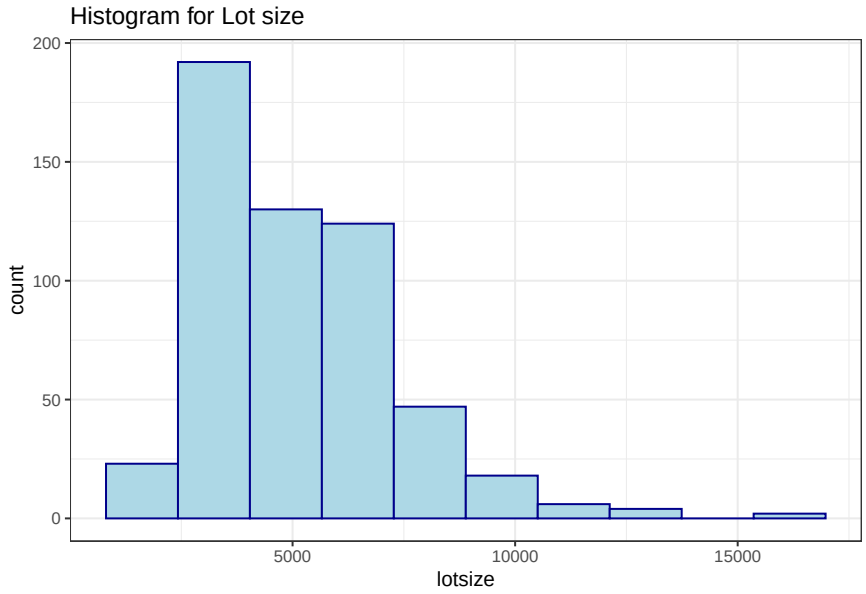
From the histogram above, the peak of the price variable occurs where the average sale prices of he houses is 50,000. The data is right skewed. Most of the sample values are clustered on the right side of the histogram. The number of houses decreases with increase in price. Most houses that were observed had low prices.

Lot size

It is a numeric variable that represents the lot size of a house in square feet. In our analysis, it will be one of the independent variables. The table below shows the summary statistics of the variable.

	Mean	Std.Dev	Min	Q1	Median	Q3	Max
lotsize	5150.27	2168.16	1650.00	3600.00	4600.00	6360.00	16200.00

We can summarize the variable graphically using an histogram

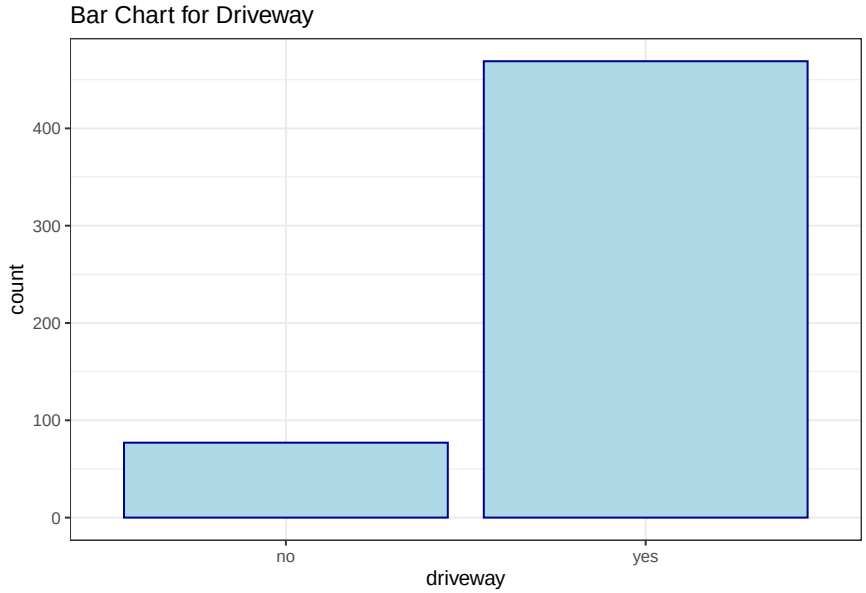


The lot size of the houses has a mean of 5150.27 and the standard deviation of 2168.16. The lot size has a right skew but has only one peak. From the position of the peak, number of houses are decreasing as the lot size is increasing. From the histogram, there are some visible outliers in the lot size.

Driveway

This is an independent categorical variable. It has two levels, `yes` and `no`. The response is *yes* if the house has a driveway, otherwise, it is *no*. In this analysis, it will be an independent variable. Below is the frequency table and the bar chart of the driveway variable.

	Freq	%	% Cum.
no	77	14.10	14.10
yes	469	85.90	100.00
Total	546	100.00	100.00

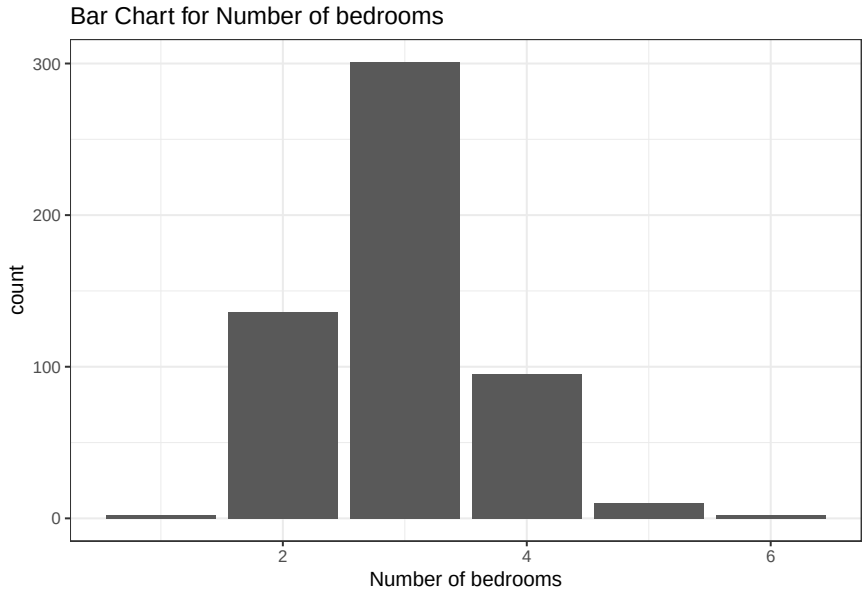


From the bar plot above, the number of houses that had a driveway were approximately 6 times the number of houses that did not have a driveway.

Number of bedrooms

The number of bedrooms is represented by *bedrooms*. It is an independent discrete variable. However, it will not be used in this analysis. Below is the summary statistics and bar chart of the number of bedrooms.

	Freq	%	% Cum.
1	2	0.37	0.37
2	136	24.91	25.27
3	301	55.13	80.40
4	95	17.40	97.80
5	10	1.83	99.63
6	2	0.37	100.00
Total	546	100.00	100.00

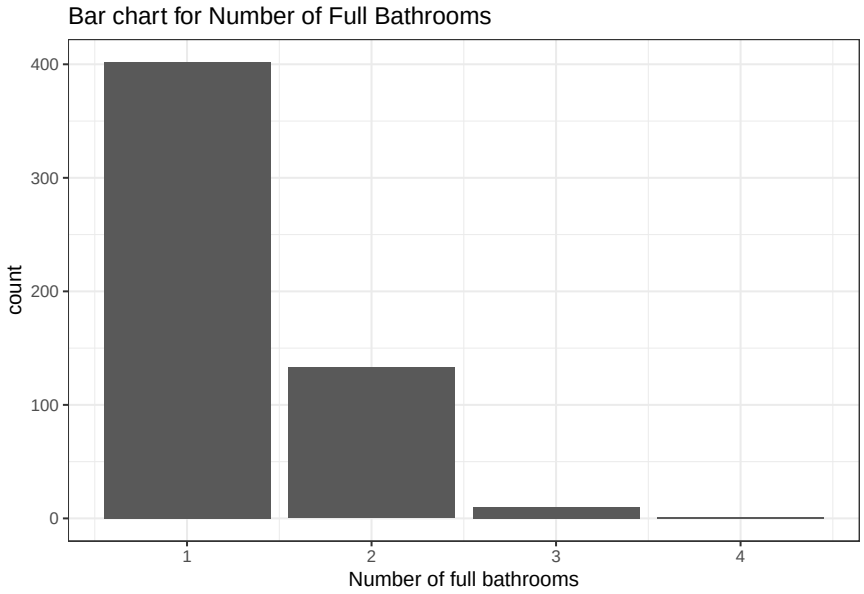


The number of bedrooms range from 1 to 6. A few houses had less than 2 bedrooms. Most houses had 2 and 3 bedrooms, and the number of bedrooms increased.

Number of full bathrooms

The number of full bathrooms is represented by *bathrms*. It is an independent discrete variable. However, it will not be used in this analysis, despite that it has an effect on the prices of houses. Below is the table of summary statistics and bar chart of the number of bathrooms.

	Freq	%	% Cum.
1	402	73.63	73.63
2	133	24.36	97.99
3	10	1.83	99.82
4	1	0.18	100.00
Total	546	100.00	100.00

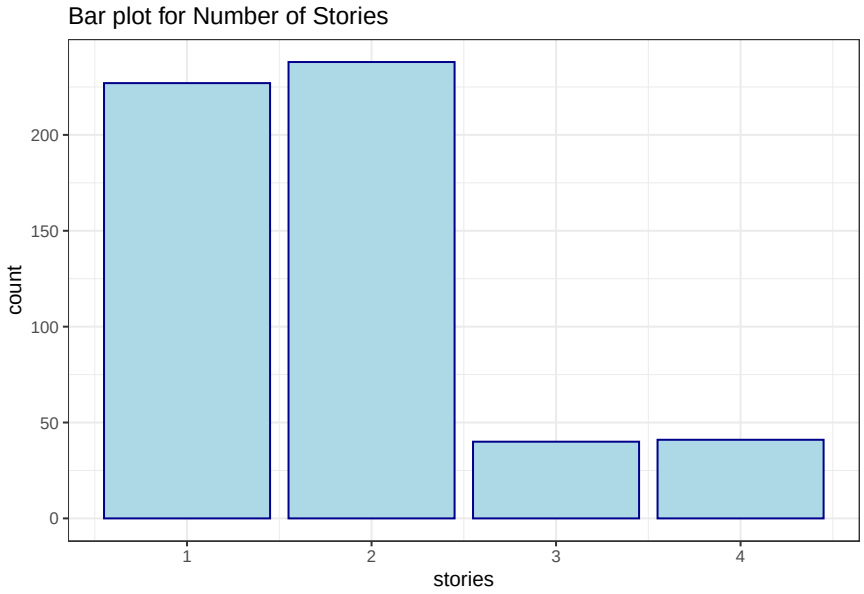


The number of bathrooms range from 1 to 4. Most houses had one bathroom the least number of houses had 4 bedrooms.

Stories

This variable contains the number of stories excluding basement. It is a discrete independent variable. This variable will not be used in our analysis. Below is the summary statistics and the bar chart of the *stories* variable.

	Freq	%	% Cum.
1	227	41.58	41.58
2	238	43.59	85.16
3	40	7.33	92.49
4	41	7.51	100.00
Total	546	100.00	100.00

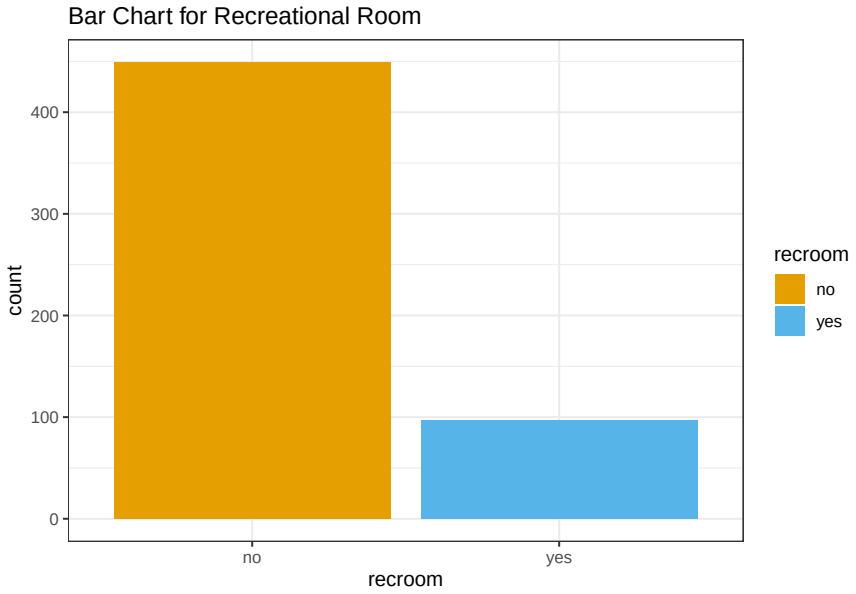


The number of stories range from 1 to 4. Most houses had 2 stories. The number of houses that had 3 and 4 stories have almost equal frequencies.

Recreational room

In the dataset, the variable is represented by `recroom`. It is an independent categorical variable. It has two levels, `yes` and `no`. The response is *yes* if the house has a recreational room, otherwise, it is *no*. The variable will not be used in this analysis. Below is the frequency table and the bar chart of this variable.

	Freq	%	% Cum.
no	449	82.23	82.23
yes	97	17.77	100.00
Total	546	100.00	100.00

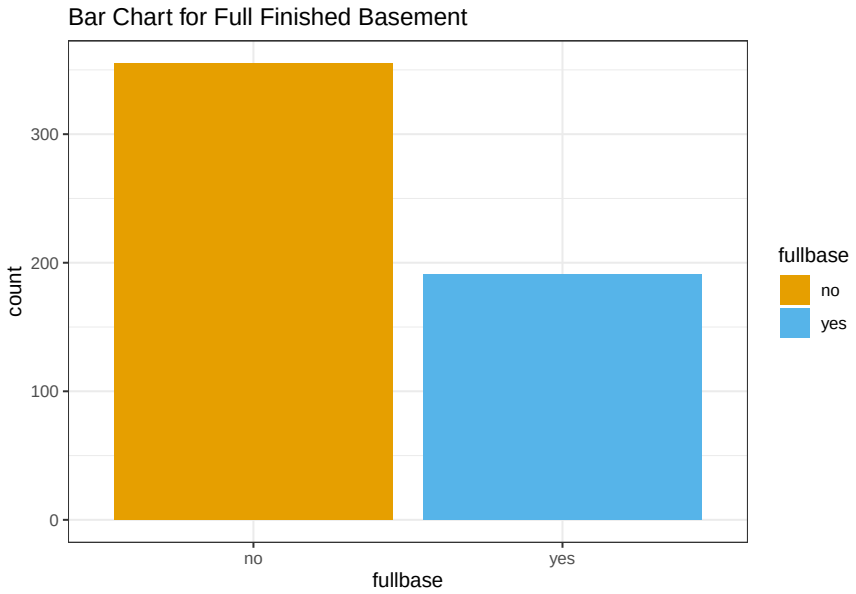


Most houses did not have a recreation room. Only a few houses had a recreation room.

Full finished basement

In the dataset, the variable is represented by `fullbase`. It is an independent categorical variable. It has two levels, `yes` and `no`. The response is *yes* if the house has a full finished basement, otherwise, it is *no*. The variable will not be used in this analysis. Below is the frequency table and the bar chart of this variable.

	Freq	%	% Cum.
no	355	65.02	65.02
yes	191	34.98	100.00
Total	546	100.00	100.00

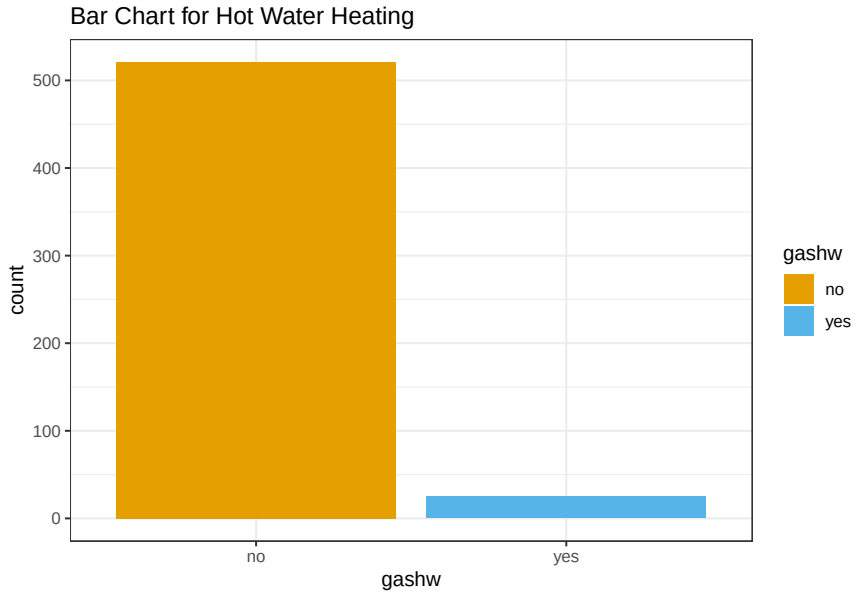


From the above bar chart, most houses had a furnished basement.

Gas for hot water heating

In the dataset, the variable is represented by `gashw`. It is an independent categorical variable. It has two levels, `yes` and `no`. The response is *yes* if the house has a gas for hot water heating, otherwise, it is *no*. The variable will not be used in this analysis. Below is the frequency table and the bar chart of this variable.

	Freq	%	% Cum.
no	521	95.42	95.42
yes	25	4.58	100.00
Total	546	100.00	100.00

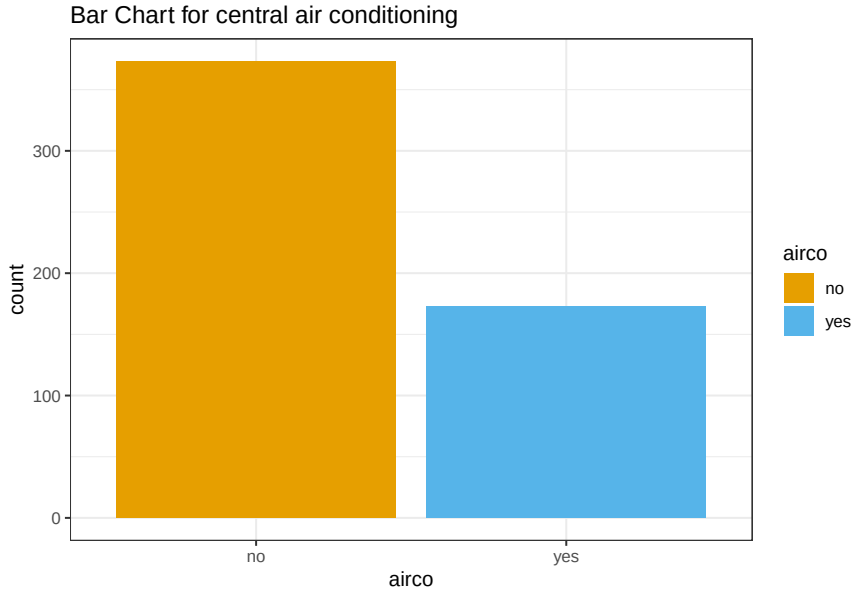


Most houses had a gas for heating water. Very few houses did not have a gas for heating water

Central air conditioning

In the dataset, the variable is represented by `airco`. It is an independent categorical variable. It has two levels, `yes` and `no`. The response is *yes* if the house has a central air conditioning, otherwise, it is *no*. The variable will not be used in this analysis. Below is the frequency table and the bar chart of this variable.

	Freq	%	% Cum.
no	373	68.32	68.32
yes	173	31.68	100.00
Total	546	100.00	100.00

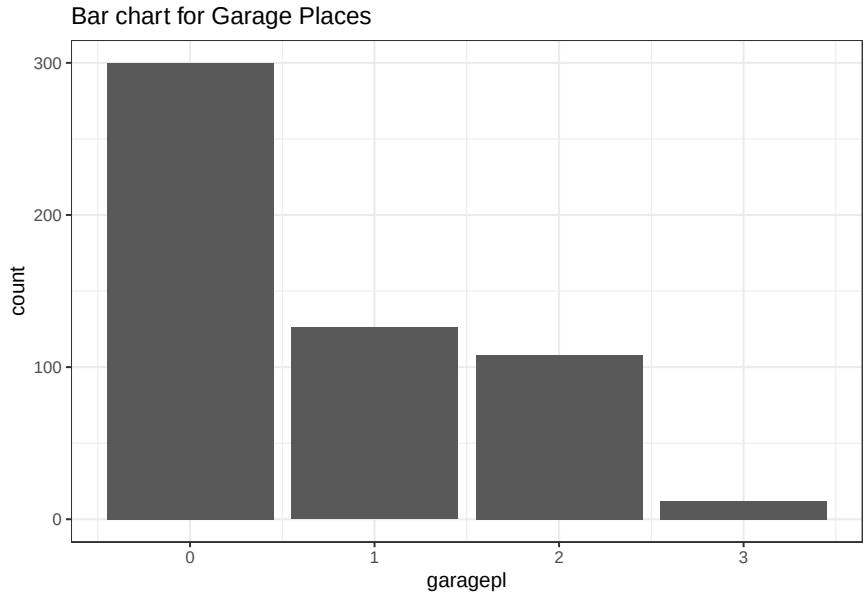


More than half of the observed houses had a central air conditioning.

Garage places

The number of garage places is represented by *garagepl*. It is an independent descrete variable. However, it will not be used in this analysis. Below is the summary statistics and histogram of the number of garage places.

	Freq	%	% Cum.
0	300	54.95	54.95
1	126	23.08	78.02
2	108	19.78	97.80
3	12	2.20	100.00
Total	546	100.00	100.00

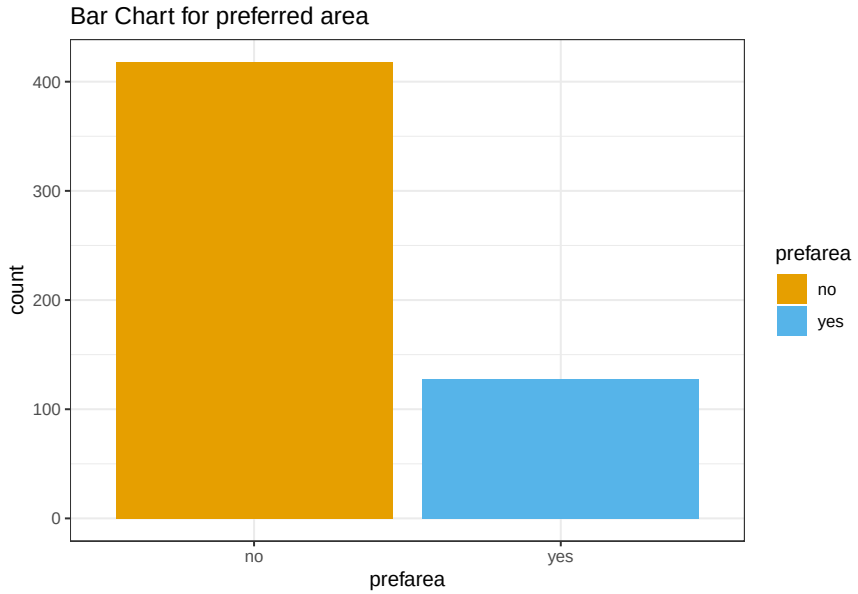


The number of garage places range from 0 to 3. Most houses did not have a garage place. The frequency decreased as the number of garage places increased.

Preferred area

In the dataset, the variable is represented by `prefarea`. It is an independent categorical variable. It has two levels, `yes` and `no`. The response is *yes* if the house located in the preferred neighborhood of the city, otherwise, it is *no*. The variable will not be used in this analysis. Below is the frequency table and the bar chart of this variable.

	Freq	%	% Cum.
no	373	68.32	68.32
yes	173	31.68	100.00
Total	546	100.00	100.00



More than half of the observed houses were not located in the preferred neighborhood of the city.

Simple Linear Regression

The response value for this analysis is price and the explanatory variable is lot size, which are both numeric. Below is the summary of the simple linear regression.

Observations	546
Dependent variable	price
Type	OLS linear regression

F(1,544)	219.056
R^2	0.287
Adj. R^2	0.286

	Est.	S.E.	t val.	p
(Intercept)	34136.192	2491.064	13.703	0.000
lotsize	6.599	0.446	14.801	0.000

Standard errors: OLS

Scatter plot

Even if not all points fall exactly on the line, it is important to describe the relationship using a line. The scatter plot below summarizes the relationship between the two variables.



From the scatter plot, the data show an uphill pattern as you move from left to right. This indicates a positive relationship between price and lot size. As the lot size increases, the price tend to increase.

Correlation coefficient.

The strength of the linear relationship can also be indicated by the correlation coefficient. The correlation coefficient between price and lot size is 0.536 . This indicates that price and lot size are moderately correlated. The correlation coefficient is positive, indicating an indicating that as lot size increases, price also increases.

Multiple R^2

It is the ratio or amount of variance in the dependent variable which can be predicted from the independent variables. R^2 is also called a "coefficient of determination". The coefficient of determination for this model is 0.287. This implies that 28.7% of variation in the estimated price can be explained by the lot size.

Model coefficients

The linear relationship between price for each house and its lot size is described using the equation:

$$\hat{price} = 34136.192 + 6.599 \times LotSize$$

The slope of the model is 6.599. It is the expected increase in price that would occur if the lot size increases by one square foot. The intercept of the linear model is 34136.192. It is the expected price of the house when the lot size is zero, which does not make sense because a house with a lot size of zero can never exist and the price should also be zero. The minimum value of lot size is 1650 square foot. Zero is outside this range and therefore we should not consider the intercept as valid. Since we do not have data collected where the lot size is less than 1650 square foot, we cannot extrapolate anything below that range.

Confidence intervals

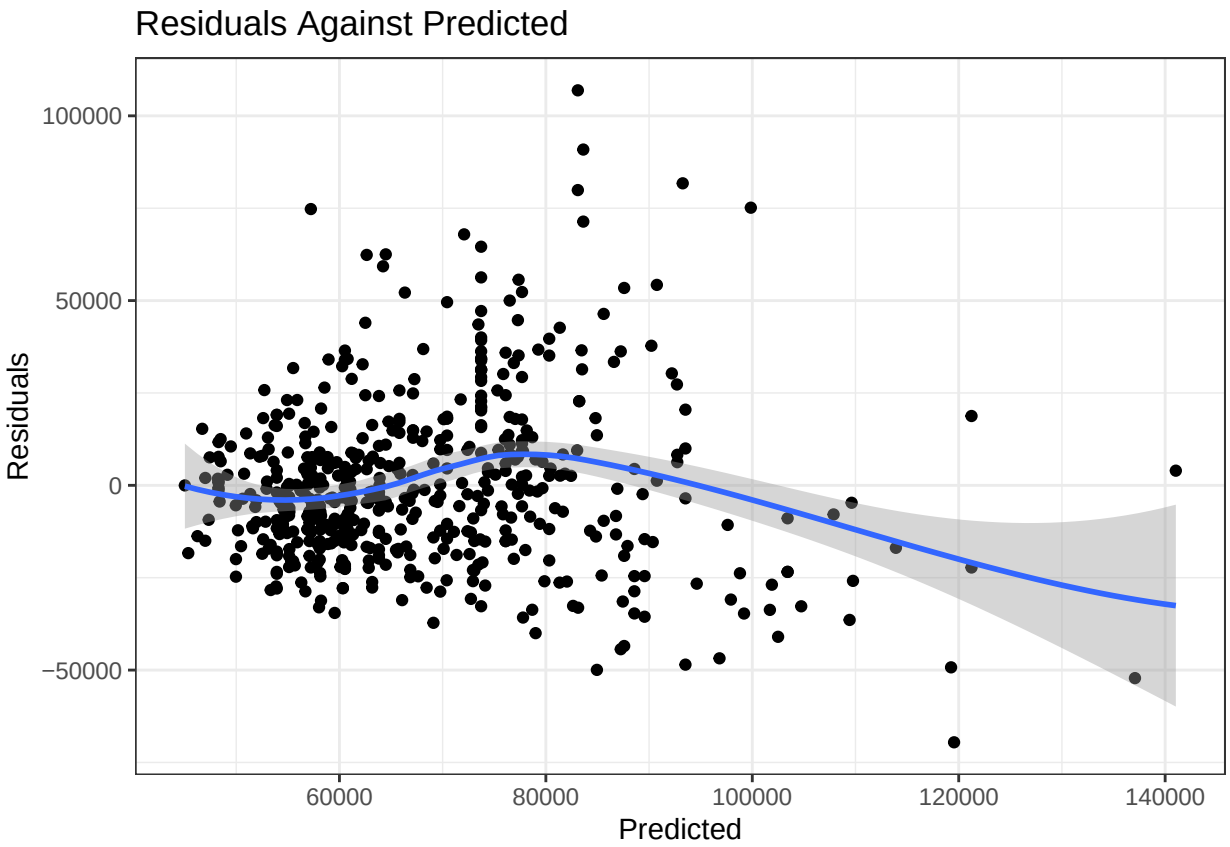
The confidence intervals are built from the estimates of the model and the standard errors. The 95% confidence intervals for both the intercept and the slope are summarized below.

```
##                    2.5 %       97.5 %
## (Intercept) 29242.909795 39029.473335
## lotsize         5.722976     7.474559
```

We are 95% confident that the slope for predicting price of a house from lot size is between 5.723 and 7.475, and the intercept is between 29242.910 and 39029.473.

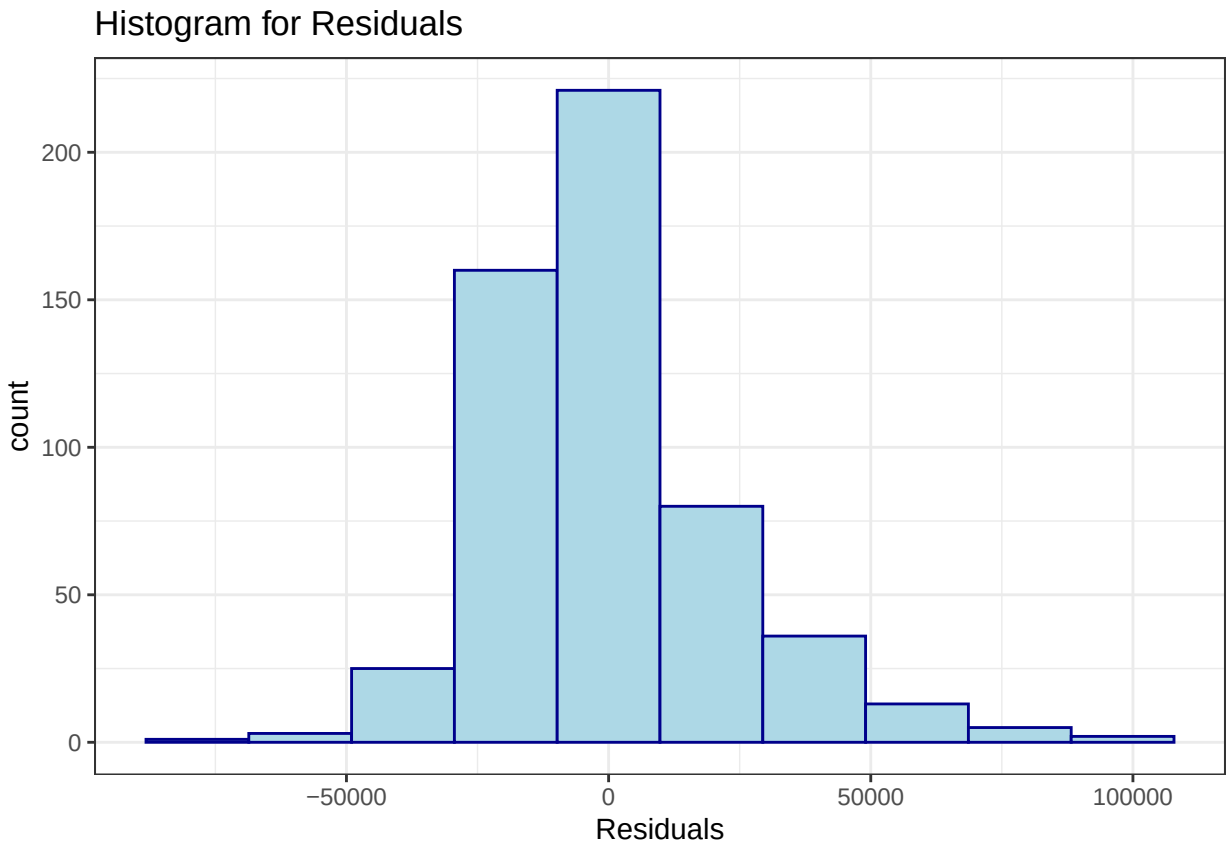
Checking assumptions

To check the assumptions of the linear model, we use the diagnostic plots. For the linearity condition, we will use a scatter plot for the residuals. The data should show a linear trend.



The data does not show a linear trend. Therefore, linear model might not be the best for this data because there are lots of negative values for very small and very large values of lot size and lot of positive values for values of lot size at the middle. However, we will proceed with caution. There appear to be a few outliers in the scatter plot. Most observations are for small values of lot size.

The nearly normal residuals condition is checked with a histogram.



The histogram is nearly normal and centered at approximately 0. There do not appear to be any outliers.

Influential points.

There are points of high leverage that invoke influence on the slope of the line. If we had fitted the line without these influential points, it could result into a stronger relationship between the variables.

Conclusion on the simple linear model

To test if there is a relationship between the price of a house and lot size of a house, we check to see if the slope is equal to zero. i.e,

$H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$.

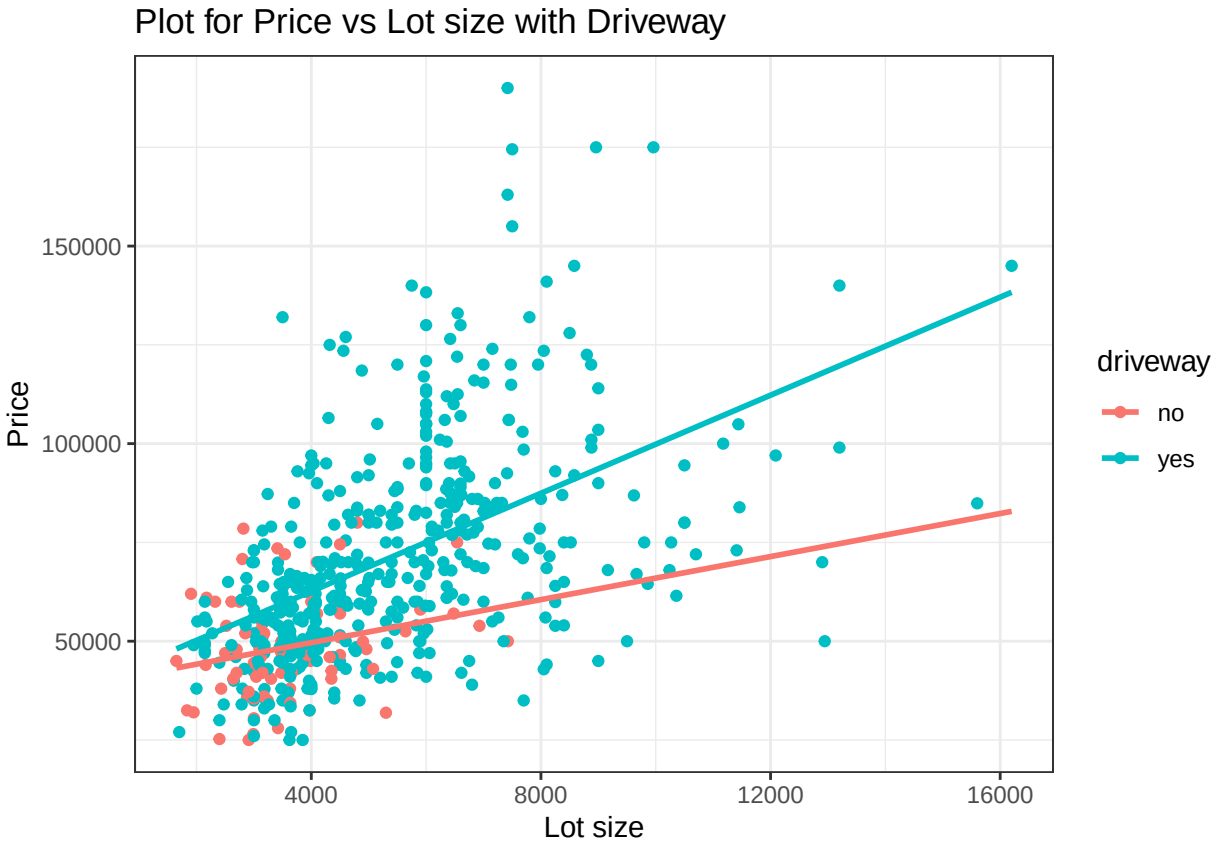
The p- value is less than the significance level ($\alpha = 0.05$). Therefore, we reject the null hypothesis in favor of the alternative and the data provide strong evidence that the true slope is different from zero. There is significant evidence of a strong positive linear relationship between price of the house and the lot size of the house. Since the null hypothesis is rejected, we can also conclude that there is statistical evidence to show that lot size of a house is a significant predictor of sale price of the house.

Adding an Indicator Variable

The price of a house can be predicted based on a categorical predictor. The indicator variable that is considered in this analysis is `Driveway`. It has two levels, `yes` and `no`.

Scatterplot

We will start by visualizing the data using a scatterplot between the sale price and lot size with respect to the driveway.



Some observations can be made from the above scatterplot. The slope of the houses with a driveway is bigger than the slope of the houses with no driveway. This indicates that the price of the houses with a driveway is expected to be higher than the prices of the houses with no driveway. Generally, driveway might be a significant predictor of price.

The houses with no driveway are highly distributed where the lot sizes are small and no distributions where the sizes are very high. Therefore, the houses with no driveway are more likely to have a small lot size.

Adjusted R^2.

Below is the summary of the model with the indicator variable.

Observations	546
Dependent variable	price
Type	OLS linear regression

F(2,543)	121.530
R^2	0.309
Adj. R^2	0.307

	Est.	S.E.	t val.	p
(Intercept)	26750.941	3026.293	8.840	0.000
lotsize	6.046	0.459	13.177	0.000
drivewayyes	11911.569	2855.605	4.171	0.000

Standard errors: OLS

As expected, the value of R^2 increases after an indicator variable is added, since R^2 will always increase even if the added predictor variable does not make sense. The adjusted R^2 is used in this case. The adjusted R^2 of the model with the indicator variable is higher than the original model with only the numeric predictor. The model with the indicator variable is the better model. Therefore, driveway is a significant predictor of price.

Categorical Variable Coefficient

The linear relationship between the three variables is explained with the equation:

$$\hat{price} = 26750.941 + 6.046 \times lotSize + 11911.569 \times yes$$

The reference level in the driveway is `no` since it did not receive any coefficient. On average, the sale price of a house is expected to increase by 11911.569 when the house has a driveway. The increase is significant because the p- value(p= 0.000) is less than the significance level ($\alpha = 0.05$). If the house does not have a driveway, the coefficient of the categorical variable is zero.

Confidence interval

Below are the 95% confidence intervals of the model with an indicator variable.

```
##                    2.5 %       97.5 %
## (Intercept) 20806.265289 32695.616918
## lotsize         5.144797     6.947365
## drivewayyes  6302.182871 17520.955944
```

We are 95% confident that the true difference in price of a house with a driveway and a house with no driveway is between 6302.183 and 17520.956.

Report

Introduction

Usually when people want to buy a house, they look for a house which has a reasonable cost, and which has all the desired features they want in the house. The house price prediction will help them to decide whether the house they desire to buy is worth of the price or not.Similar is the case with people who want to sell the house. By making use of the house price prediction system, the seller would be able to decide what all features he/she could add in the house so that the house can be sold for a higher price.

The objective of this analysis is predicting house prices on the basis of various parameters such as lot size and availability of driveway.This will allow the buyer to get an idea of what amount of money he/she has to spend in order to buy the desired house. It will also allow the seller to get information regarding what is the house's real worth and how he/she can maximize the profit gained by selling the house.

Dataset Gathering

The dataset used in the analysis contains 546 number of observations and 3 variables. Out of the 3 variables, 2 are numeric and 1 is categorical. According to the dataset, the prices are given in USD and the plot size has been given in square feet. The price column in the dataset is the dependent variable and the rest of the columns are independent variables.

```
##   price lotsize driveway
## 1 42000    5850      yes
## 2 38500    4000      yes
## 3 49500    3060      yes
## 4 60500    6650      yes
## 5 61000    6360      yes
## 6 66000    4160      yes
```

Exploratory Analysis and Data Summary

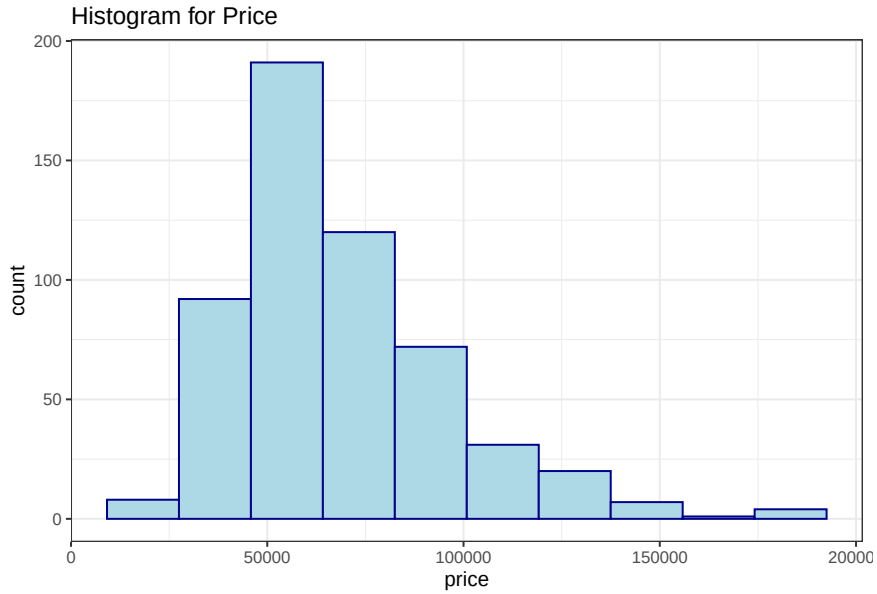
This section contains the summary of all the variables that were used in the analysis.

Price

It is a numeric continuous variable that represents the sale price of a house. It is a dependent variable. The table below shows the summary statistics of the variable.

	Mean	Std.Dev	Min	Q1	Median	Q3	Max
price	68121.60	26702.67	25000.00	49000.00	62000.00	82000.00	190000.00

We can summarize the distribution of the `price` variable using a histogram.



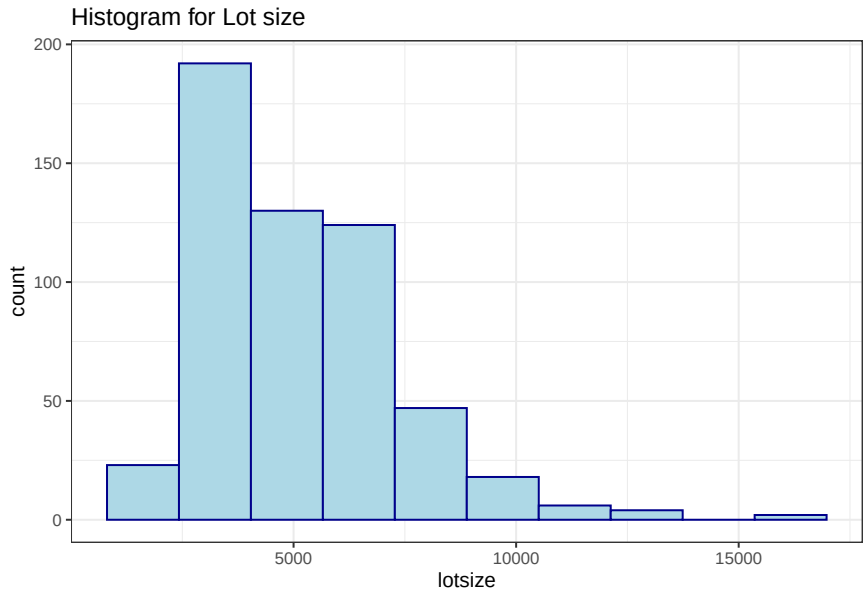
From the histogram above, the peak of the price variable occurs where the average sale prices of he houses is 50,000. The data is right skewed. Most of the sample values are clustered on the right side of the histogram. The number of houses decreases with increase in price. Most houses that were observed had low prices.

Lot size

It is a numeric variable that represents the lot size of a house in square feet. It is an independent variable. The table below shows the summary statistics of the variable.

	Mean	Std.Dev	Min	Q1	Median	Q3	Max
lotsize	5150.27	2168.16	1650.00	3600.00	4600.00	6360.00	16200.00

We can summarize the variable graphically using an histogram

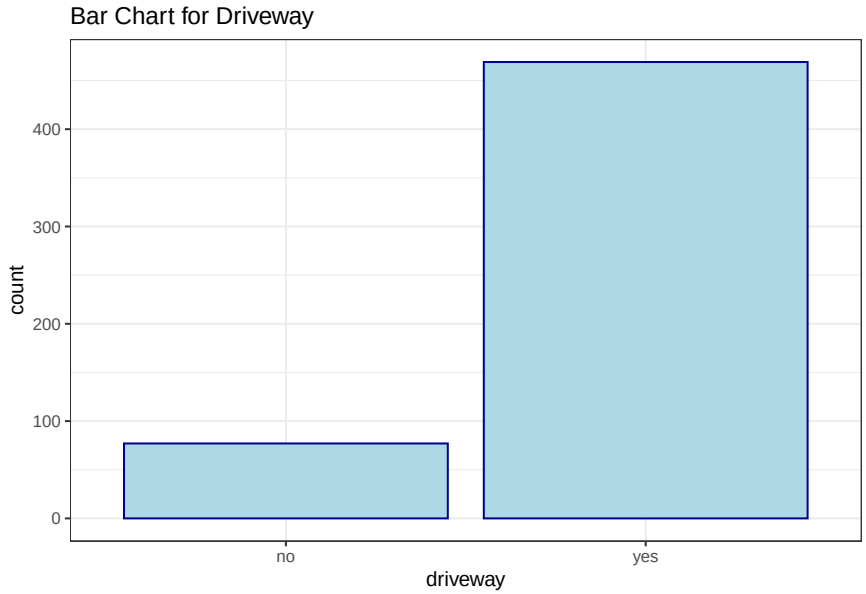


The lot size of the houses has a mean of 5150.27 and the standard deviation of 2168.16. The lot size has a right skew but has only one peak. From the position of the peak, number of houses are decreasing as the lot size is increasing. From the histogram, there are some visible outliers in the lot size.

Driveway

This is an independent categorical variable. It has two levels, `yes` and `no`. The response is *yes* if the house has a driveway, otherwise, it is *no*. In this analysis, it will be an independent variable. Below is the frequency table and the bar chart of the driveway variable.

	Freq	%	% Cum.
no	77	14.10	14.10
yes	469	85.90	100.00
Total	546	100.00	100.00



From the bar plot above, the number of houses that had a driveway were approximately 6 times the number of houses that did not have a driveway.

Simple Linear Regression

The first step is comparing price with the lot size. A simple linear regression was conducted and below is the summary of the output.

Observations	546
Dependent variable	price
Type	OLS linear regression

F(1,544)	219.056
R^2	0.287
Adj. R^2	0.286

	Est.	S.E.	t val.	p
(Intercept)	34136.192	2491.064	13.703	0.000
lotsize	6.599	0.446	14.801	0.000

Standard errors: OLS

Even if not all points fall exactly on the line, it is important to describe the relationship using a line. The scatter plot below summarizes the relationship between the two variables.



From the scatter plot, the data show an uphill pattern as you move from left to right. This indicates a positive relationship between price and lot size. As the lot size increases, the price tend to increase.

The strength of the linear relationship can also be indicated by the correlation coefficient. The correlation coefficient between price and lot size is 0.536 . This indicates that price and lot size are moderately correlated. The correlation coefficient is positive, indicating an indicating that as lot size increases, price also increases.

Multiple R^2 is the ratio or amount of variance in the dependent variable which can be predicted from the independent variables. R^2 is also called a "coefficient of determination". The coefficient of determination for this model is 0.287. This implies that 28.7% of variation in the estimated price can be explained by the lot size.

The linear relationship between price for each house and its lot size is described using the equation:

$$\hat{price} = 34136.192 + 6.599 \times LotSize$$

The slope of the model is 6.599. It is the expected increase in price that would occur if the lot size increases by one square foot. The intercept of the linear model is 34136.192. It is the expected price of the house when the lot size is zero, which does not make sense because a house with a lot size of zero can never exist and the price should also be zero. The minimum value of lot size is 1650 square foot. Zero is outside this range and therefore we should not consider the intercept as valid. Since we do not have data collected where the lot size is less than 1650 square foot, we cannot extrapolate anything below that range.

The confidence intervals are built from the estimates of the model and the standard errors. The 95% confidence intervals for both the intercept and the slope are summarized below.

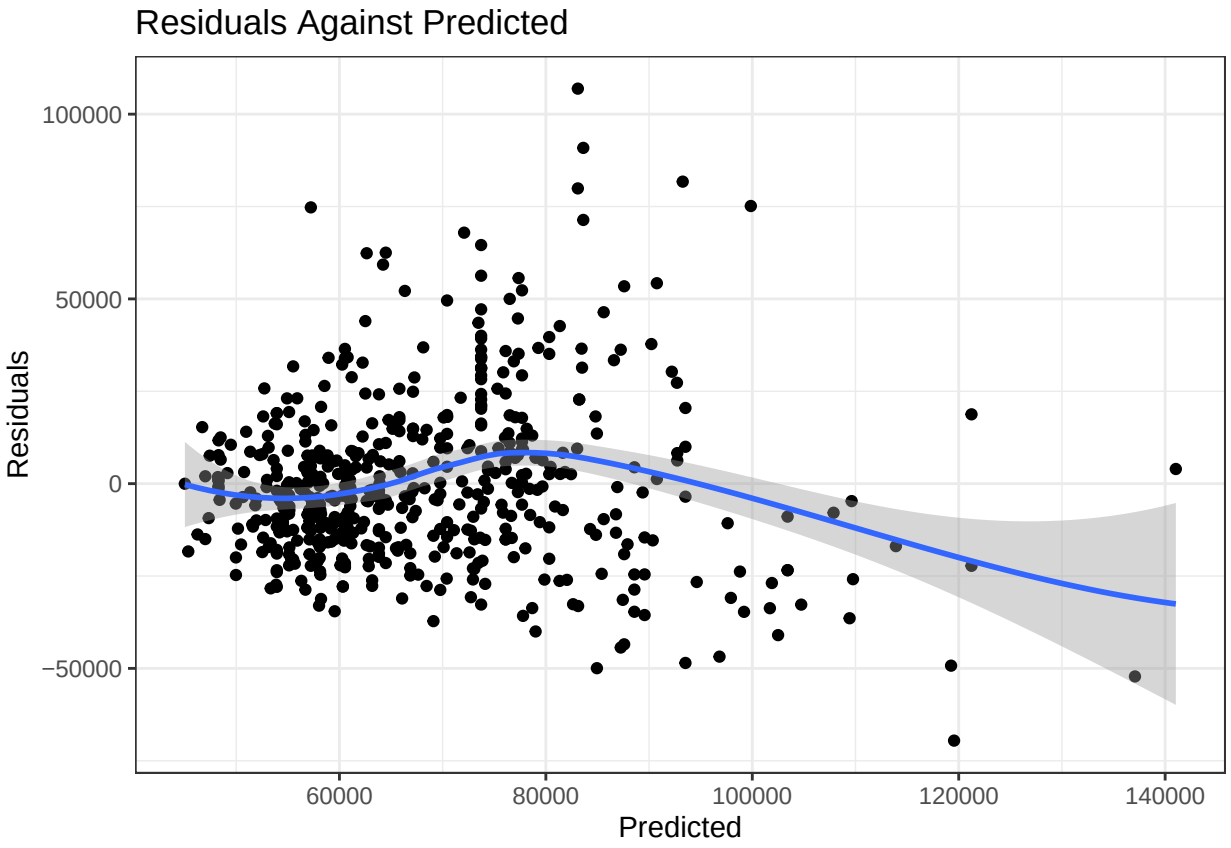
```
##                   2.5 %       97.5 %
## (Intercept) 29242.909795 39029.473335
```

```
## lotsize          5.722976     7.474559
```

We are 95% confident that the slope for predicting price of a house from lot size is between 5.723 and 7.475, and the intercept is between 29242.910 and 39029.473.

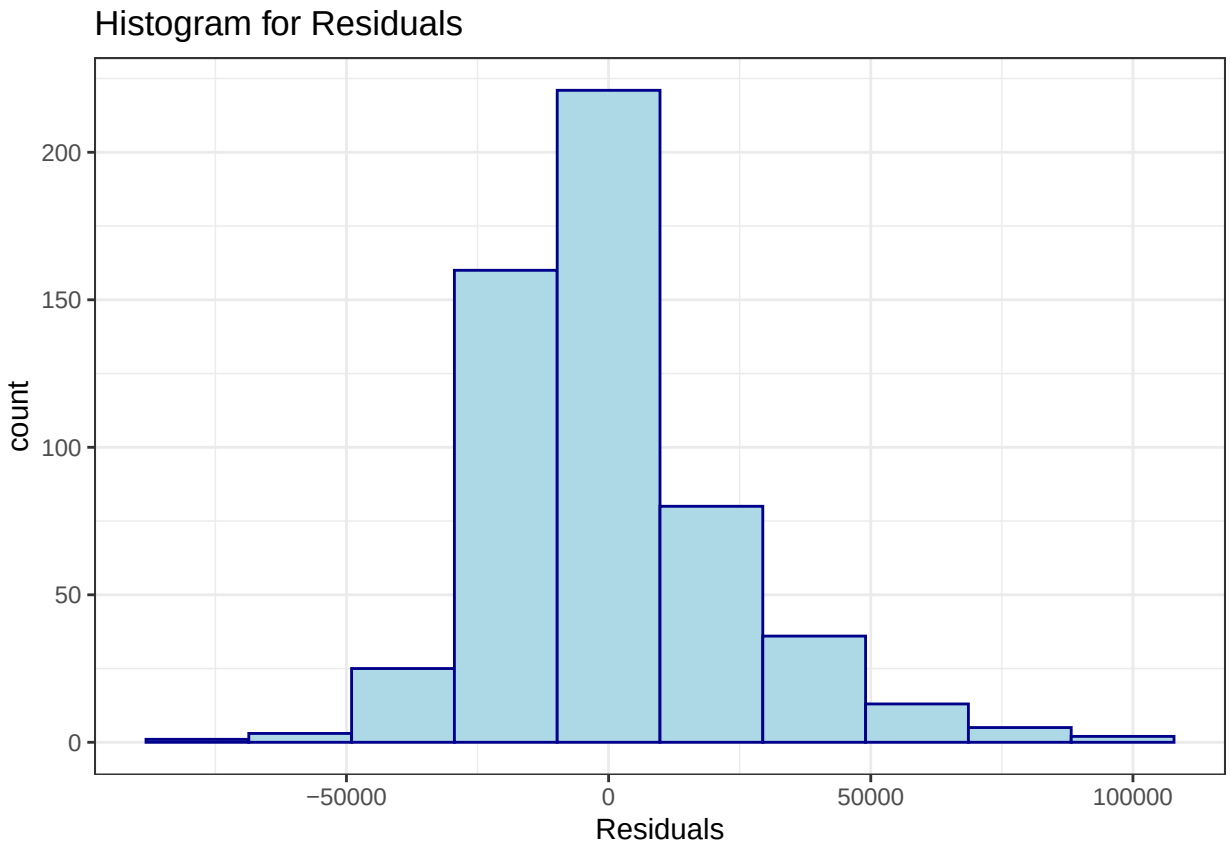
Checking assumptions

To check the assumptions of the linear model, we use the diagnostic plots. For the linearity condition, we will use a scatter plot for the residuals. The data should show a linear trend.



The data does not show a linear trend. Therefore, linear model might not be the best for this data because there are lots of negative values for very small and very large values of lot size and lot of positive values for values of lot size at the middle. However, we will proceed with caution. There appear to be a few outliers in the scatter plot. Most observations are for small values of lot size.

The nearly normal residuals condition is checked with a histogram.



The histogram is nearly normal and centered at approximately 0. There do not appear to be any outliers.

From the scarrer plot, there are points of high leverage that invoke influence on the slope of the line. If we had fitted the line without these influential points, it could result into a stronger relationship between the variables.

Conclusion on the simple linear model

To test if there is a relationship between the price of a house and lot size of a house, we check to see if the slope is equal to zero. i.e,

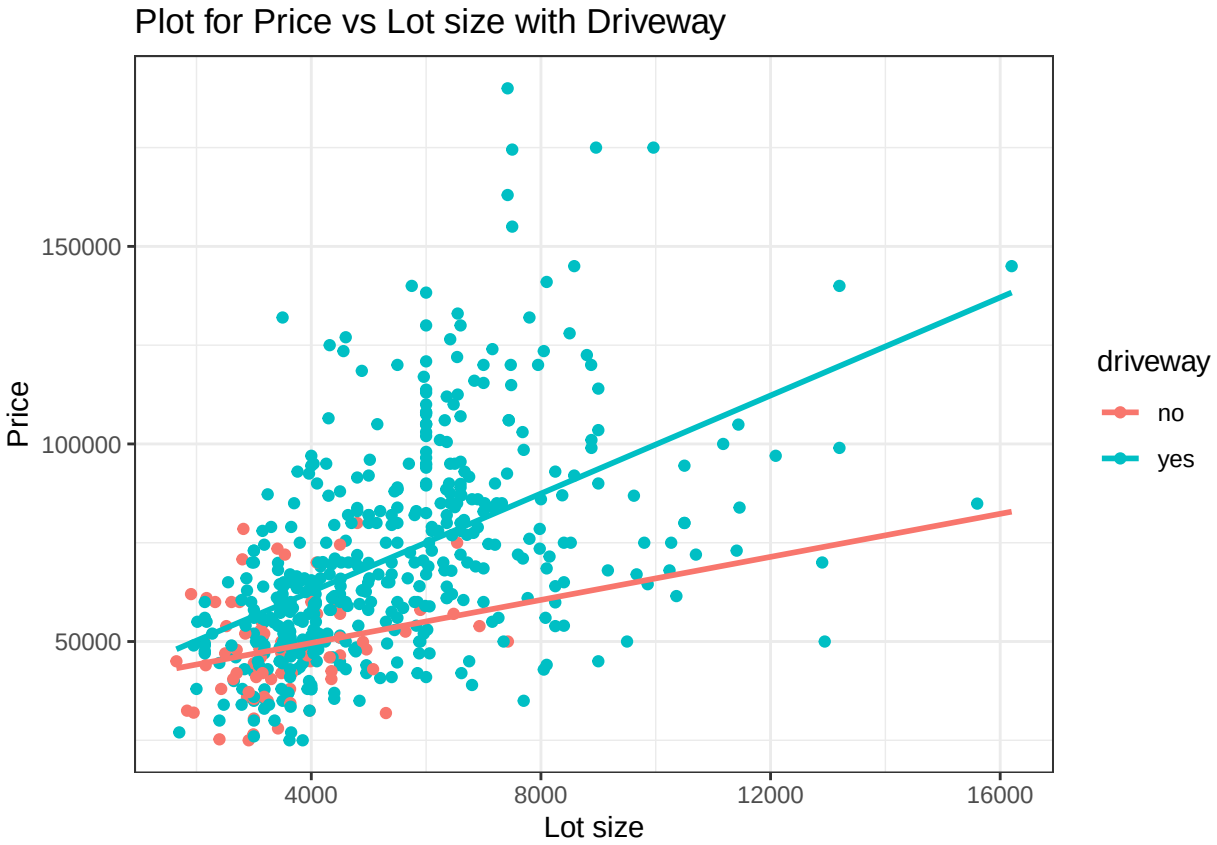
$H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$.

The p- value is less than the significance level ($\alpha = 0.05$). Therefore, we reject the null hypothesis in favor of the alternative and the data provide strong evidence that the true slope is different from zero. There is significant evidence of a strong positive linear relationship between price of the house and the lot size of the house. Since the null hypothesis is rejected, we can also conclude that there is statistical evidence to show that lot size of a house is a significant predictor of sale price of the house.

Adding an Indicator Variable

The price of a house can be predicted based on a categorical predictor. The indicator variable that is considered in this analysis is `Driveway`. It has two levels, `yes` and `no`.

I started by visualizing the data using a scatterplot between the sale price and lot size with respect to the driveway.



Some observations can be made from the above scatterplot. The slope of the houses with a driveway is bigger than the slope of the houses with no driveway. This indicates that the price of the houses with a driveway is expected to be higher than the prices of the houses with no driveway. Generally, driveway might be a significant predictor of price.

The houses with no driveway are highly distributed where the lot sizes are small and no distributions where the sizes are very high. Therefore, the houses with no driveway are more likely to have a small lot size.

Below is the summary of the model with the indicator variable.

Observations	546
Dependent variable	price
Type	OLS linear regression

F(2,543)	121.530
R^2	0.309
Adj. R^2	0.307

	Est.	S.E.	t val.	p
(Intercept)	26750.941	3026.293	8.840	0.000
lotsize	6.046	0.459	13.177	0.000
drivewayyes	11911.569	2855.605	4.171	0.000

Standard errors: OLS

As expected, the value of R^2 increases after an indicator variable is added, since R^2 will always increase even

if the added predictor variable does not make sense. The adjusted R^2 is used in this case. The adjusted R^2 of the model with the indicator variable is higher than the original model with only the numeric predictor. The model with the indicator variable is the better model. Therefore, driveway is a significant predictor of price.

The linear relationship between the three variables is explained with the equation:

$$\hat{price} = 26750.941 + 6.046 \times lotSize + 11911.569 \times yes$$

The reference level in the driveway is `no` since it did not receive any coefficient. On average, the sale price of a house is expected to increase by 11911.569 when the house has a driveway. The increase is significant because the p- value(p= 0.000) is less than the significance level ($\alpha = 0.05$). If the house does not have a driveway, the coefficient of the categorical variable is zero.

Below are the 95% confidence intervals of the model with an indicator variable.

```
##                     2.5 %       97.5 %
## (Intercept) 20806.265289 32695.616918
## lotsize         5.144797     6.947365
## drivewayyes  6302.182871 17520.955944
```

From this, we can conclude that we are 95% confident that the true difference in price of a house with a driveway and a house with no driveway is between 6302.183 and 17520.956.

Conclusion

In this analysis, I used the regression model to predict the price of different houses. All the steps required for the successful completion of the house price prediction system have been completed. It is seen that the multiple linear regression model with the indicator variable is suitable for the purpose of predicting house prices.

Future Scope

There are some improvements and additions which can be done. Increasing and updating the dataset on a regular basis. This will make the prediction system more correct and accurate.