Name: **Varun Sudhir**

Reg No: **21BDS0040**

Exploratory Data Analysis Assignment – 4

First,we will import the dataset and display it . We are working with a dataset that describes the pertinent features with regards to pricing of houses in the USA

```
# Varun Sudhir 21BDS0040
dataset <- read.csv("C:/Users/Varun/Desktop/USA_Housing.csv")
View(dataset)
```

	Avg..Area.Income	Avg..Area.House.Age	Avg..Area.Number.of.Rooms	Avg..Area.Number.of.Bedrooms	Area.Population	Price	Address
1	79545.46	5.682861	7.009188	4.09	23086.801	1059033.6	208 Michael Ferry Apt. 674 Laurabury, NE 37010-5101
2	79248.64	6.002900	6.730821	3.09	40173.072	1505890.9	188 Johnson Views Suite 079 Lake Kathleen, CA 48958
3	61287.07	5.865890	8.512727	5.13	36882.159	1058988.0	9127 Elizabeth Stravenue Danieltown, WI 06482-3489
4	63345.24	7.188236	5.586729	3.26	34310.243	1260616.8	USS Barnett FPO AP 44820
5	59982.20	5.040555	7.839388	4.23	26354.109	630943.5	USNS Raymond FPO AE 09386
6	80175.75	4.988408	6.104512	4.04	26748.428	1068138.1	06039 Jennifer Islands Apt. 443 Tracyport, KS 16077
7	64698.46	6.025336	8.147760	3.41	60828.249	1502055.8	4759 Daniel Shoals Suite 442 Nguyenburgh, CO 20247
8	78394.34	6.989780	6.620478	2.42	36516.359	1573936.6	972 Joyce Viaduct Lake William, TN 17778-6483
9	59927.66	5.362126	6.393121	2.30	29387.396	798869.5	USS Gilbert FPO AA 20957
10	81885.93	4.423672	8.167688	6.10	40149.966	1545154.8	Unit 9446 Box 0958 DPO AE 97025
11	80527.47	8.093513	5.042747	4.10	47224.360	1707045.7	6368 John Motorway Suite 700 Janetbury, NM 26854
12	50593.70	4.496513	7.467627	4.49	34343.992	663732.4	911 Castillo Park Apt. 717 Davisborough, PW 78603
13	39033.81	7.671755	7.250029	3.10	39220.361	1042814.1	209 Natasha Stream Suite 961 Huffmanland, NE 52457
14	73163.66	6.919535	5.993188	2.27	32326.123	1291331.5	829 Welch Track Apt. 992 North John, AR 26532-5136
15	69391.38	5.344776	8.406418	4.37	35521.294	1402818.2	PSC 5330, Box 4420 APO AP 08302
16	73091.87	5.443156	8.517513	4.01	23929.524	1306674.7	2278 Shannon View North Carriemouth, NM 84617
17	79706.96	5.067890	8.219771	3.12	39717.814	1556786.6	064 Hayley Unions Nicholsborough, HI 44161-1887
18	61929.08	4.788550	5.097010	4.30	24595.901	528485.2	5498 Rachel Locks New Gregoryshire, PW 54755
19	63508.19	5.947165	7.187774	5.12	35719.653	1019425.9	Unit 7424 Box 2786 DPO AE 71255
20	62085.28	5.739411	7.091808	5.49	44922.107	1030591.4	19696 Benjamin Cape Stephentown, ME 36952-4733
21	86295.00	6.627457	8.011898	4.07	47560.775	2146925.3	030 Larry Park Suite 665 Thomashaven, HI 87941-5197
22	60835.09	5.551222	6.517175	2.10	45574.742	929247.6	USNS Brown FPO AP 85833
23	64490.65	4.210323	5.478088	4.31	40358.960	718887.2	95198 Ortiz Key Port Sara, TN 24541-2855
24	60697.35	6.170484	7.150537	6.34	28140.967	743999.8	9003 Jay Plains Suite 838 Lake Elizabeth, IN 90622-0804
25	59748.86	5.339340	7.748682	4.23	27809.987	895737.1	24282 Paul Valley West Perry, MI 03169-5806
26	56974.48	8.287562	7.312880	4.33	40694.870	1453974.5	61938 Brady Falls Lewisfort, DE 61227

Let's perform statistical operations on the columns of the dataset

```
# Varun Sudhir 21BDS0040
# Statistical operations on the dataset

# Average price of the houses
mean_prices <- mean(dataset$Price)
print("The average price of the houses")
print(mean_prices)
print("Varun Sudhir 21BDS0040")
```

Output:

```
> # Average price of the houses
> mean_prices <- mean(dataset$Price)
> print("The average price of the houses")
[1] "The average price of the houses"
> print(mean_prices)
[1] 1232073
> print("Varun Sudhir 21BDS0040")
[1] "Varun Sudhir 21BDS0040"
```

```
# Calculating the variance in the Avg Area Income
var_income <- var(dataset$Avg..Area.Income)
print("The variation in the Avg Area Income is ")
print(var_income)
print("Varun Sudhir 21BDS0040")
```

Output:

```
> # Calculating the variance in the Avg Area Income
> var_income <- var(dataset$Avg..Area.Income)
> print("The variation in the Avg Area Income is ")
[1] "The variation in the Avg Area Income is "
> print(var_income)
[1] 113592777
> print("Varun Sudhir 21BDS0040")
[1] "Varun Sudhir 21BDS0040"
>
```

```
# Calculating the variance in the Avg Area Income
var_income <- var(dataset$Avg..Area.Income)
print("The variation in the Avg Area Income is ")
print(var_income)
print("Varun Sudhir 21BDS0040")
```

Output:

```
> # Calculating the standard deviation in the Area Population
> std_dev_population <- sd(dataset$Area.Population)
> print("The standard deviation in the Area Population")
[1] "The standard deviation in the Area Population"
> print(std_dev_population)
[1] 9925.65
> print("Varun Sudhir 21BDS0040")
[1] "Varun Sudhir 21BDS0040"
>
```

```
# Custom function to calculate mode
mode_custom <- function(x) {
  uniq_values <- unique(x)
  uniq_values[which.max(tabulate(match(x, uniq_values)))]
}

# Mode of Area Population
mode_value <- mode_custom(dataset$Area.Population)
print("Mode of Avg Area Number of Bedrooms")
print(mode_value)
print("Varun Sudhir 21BDS0040")
```

Output:

```
> print("Mode of Avg Area Number of Bedrooms")
[1] "Mode of Avg Area Number of Bedrooms"
> print(mode_value)
[1] 23086.8
> print("Varun Sudhir 21BDS0040")
[1] "Varun Sudhir 21BDS0040"
>
```

```
# Calculating the mode of Avg Area Income using the 'modeest' package
mode_value_mfv <- mfv(dataset$Avg..Area.Income)
print("The mode of AvgAreaIncome using the modeest package")
print(mode_value_mfv[1])
print("Varun Sudhir 21BDS0040")
```

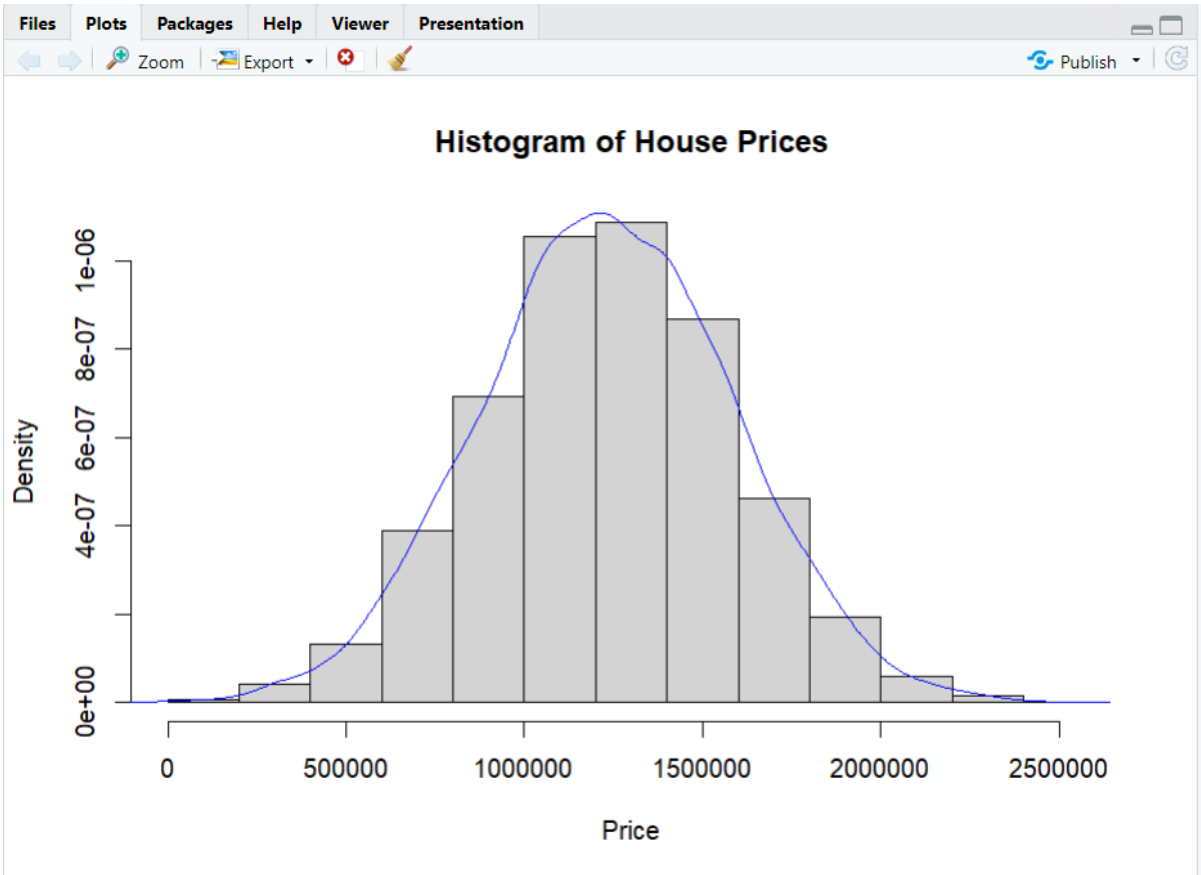
Output:

```
> print(mode_value_mfv[1])
[1] 17796.63
> print("Varun Sudhir 21BDS0040")
[1] "Varun Sudhir 21BDS0040"
>
```

Since the mfv functions returns a vector in case there are multiple modes for the particular column, we have displayed only the first element of the vector returned.

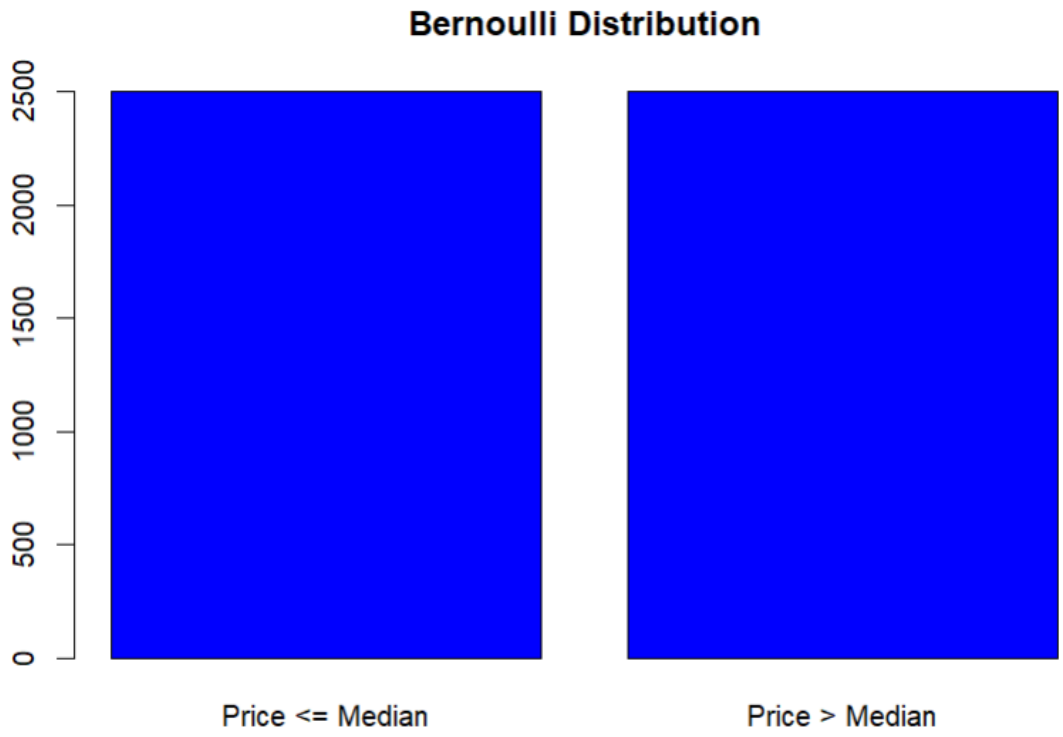
```
# Plotting a histogram to understand the variation in House Prices
# Varun Sudhir 21BDS0040
# Histogram with density overlay for 'Price'

hist(dataset$Price, probability = TRUE, main = "Histogram of House Prices",
xlab = "Price")
lines(density(dataset$Price), col = "blue")
```

Output:

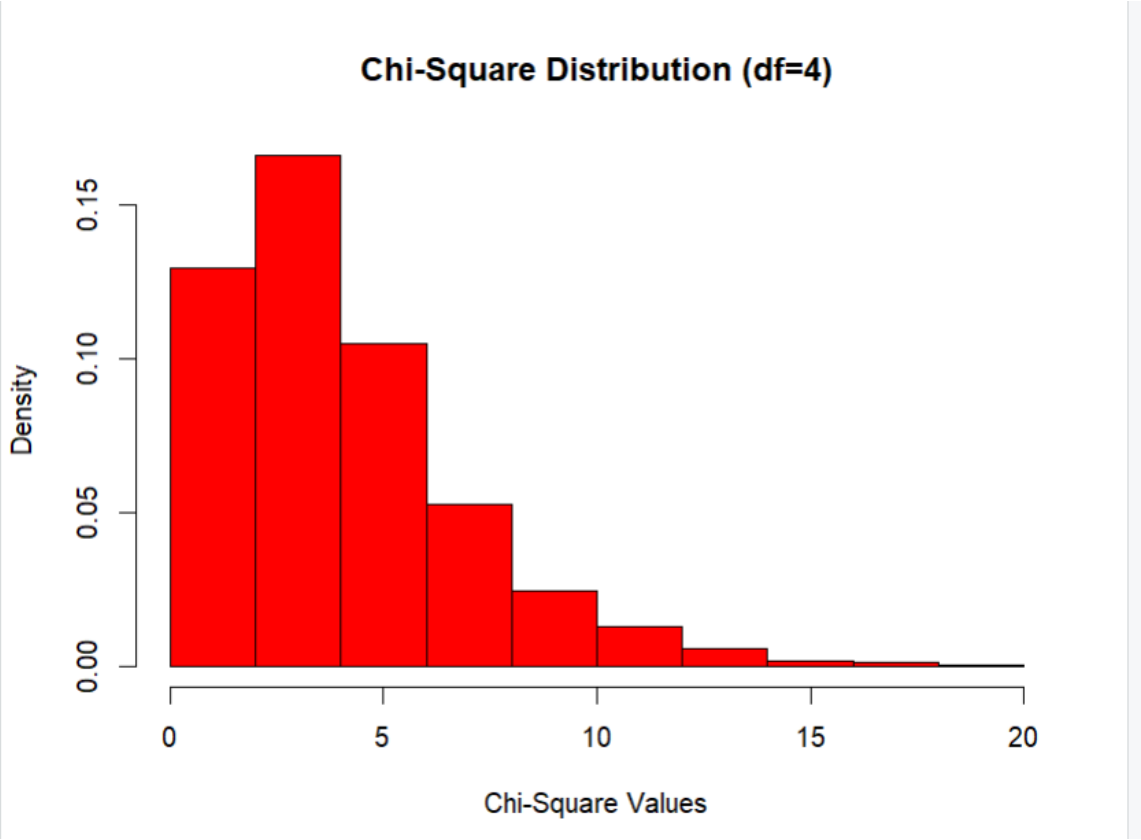


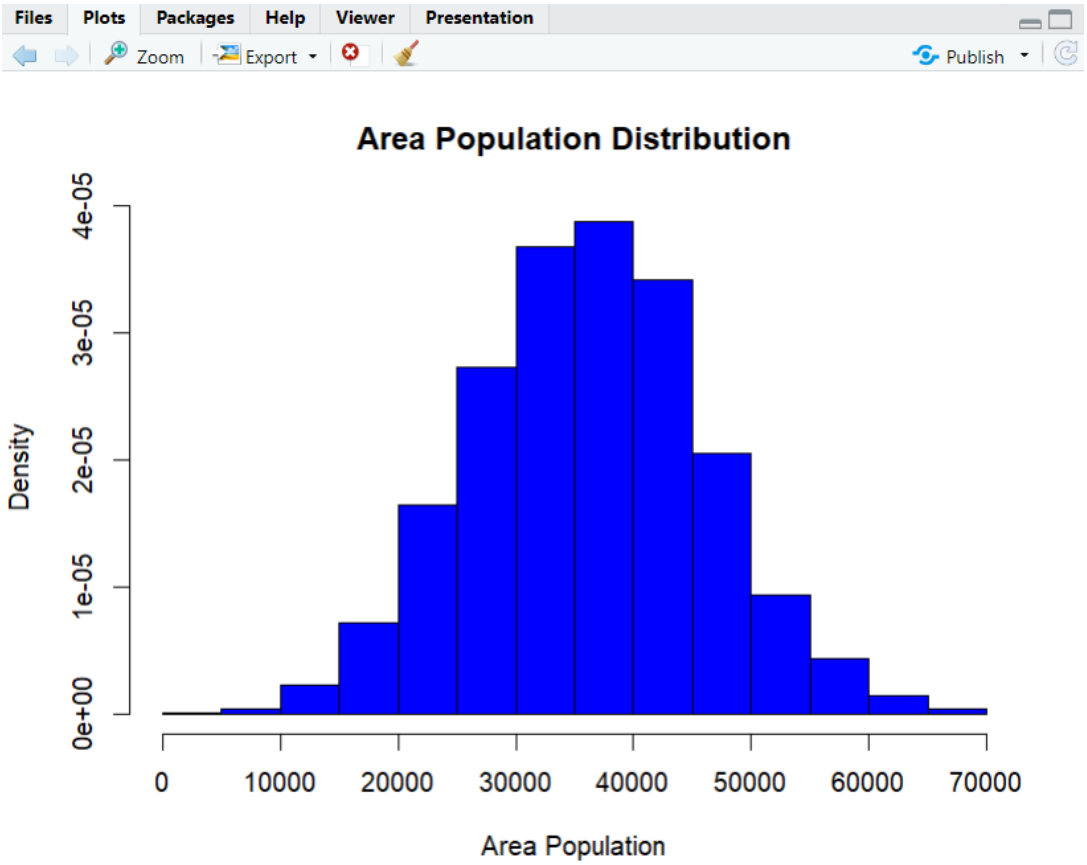
```r
# Varun Sudhir 21BDS0040
# We want to model a situation where a house price (Price) is either above or
below the median price.
# We'll treat above the median as a "success" (1) and below as a "failure"
(0).

median_price <- median(dataset$Price)
bernoulli_trials <- ifelse(dataset$Price > median_price, 1, 0)

# Calculate the probability of success (price > median)
p_success <- mean(bernoulli_trials)
cat("Probability of success (Price > median):", p_success, "\n")

# Plot the Bernoulli distribution
barplot(table(bernoulli_trials), main="Bernoulli Distribution",
names.arg=c("Price <= Median", "Price > Median"), col="blue")
```

Output:



```r
# Varun Sudhir 21BDS0040

# Set the degrees of freedom for the Chi-Square distribution
df <- 4

# Generate a Chi-Square distribution with the same number of observations as
in the dataset
chi_square_values <- rchisq(n = length(dataset$Area.Population), df = df)

# Plot Chi-Square Distribution separately
hist(chi_square_values, probability = TRUE, col = "red",
     main = "Chi-Square Distribution (df=4)",
     xlab = "Chi-Square Values")

# Plot Area Population Distribution separately
hist(dataset$Area.Population, probability = TRUE, col = "blue",
     main = "Area Population Distribution",
     xlab = "Area Population")
```






```
# Scatter Plot comparing AreaPopulation vs Price
# Varun Sudhir 21BDS0040
plot(dataset$Area.Population, dataset$Price, main="Scatter Plot of Income vs.
Price", xlab="Avg. Area Income", ylab="Price", col="green")
```

Output: