Final Written Project Report : Housing

Aakash Bhandari (300340504) Shadab Ahmed (300340441)

Summary: Here our main finding is how the Sales price of house is getting influenced by different factors(variables). And We took SalePrice as the primary variable which basically represents the price of a particular house, While all the other variables (i.e- Id, Neighborhood, HouseStyle, Heating, CentralAir, Garagears, and GarageArea) treated as secondary.

Data Cleaning: Here we only filter our desired columns (i.e- Id, Neighborhood, HouseStyle, Heating, CentralAir, Garagears, GarageArea and SalePrice) and put into a new csv file (i.e- HousingPre2.csv). Here we don't need to clean our data because the variables we chose were already cleaned, there were no NA or any kind of missing values in it.

1 Analysis

1.1 Single-Variable Analysis

1. Annalysis of Sales Price of house:

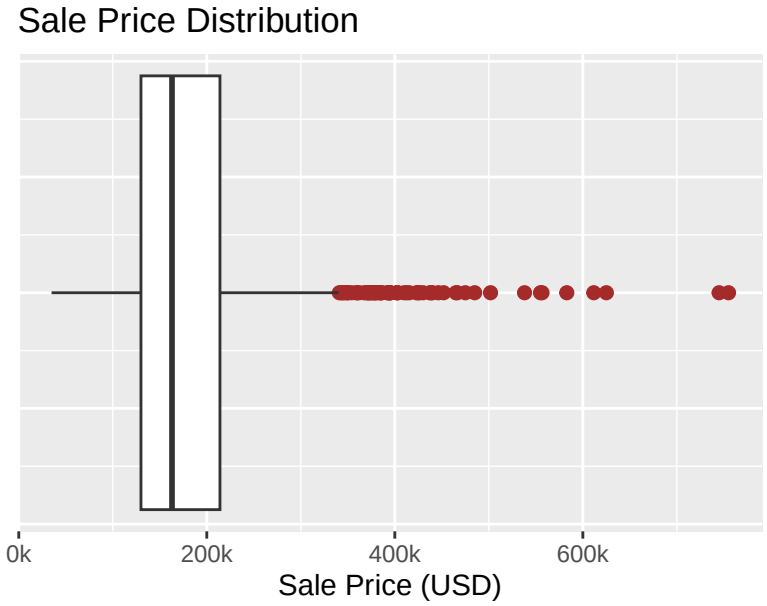
```
library(dplyr)
library(ggplot2)
library(scales)

housing <- read.csv(file.choose())
summary(housing$SalePrice)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   34900  129975  163000  180921  214000  755000
```

```
ggplot(housing,aes(x=SalePrice))+
  labs(x="Sale Price (USD)",y="")+
  ggtitle("Sale Price Distribution")+
  geom_boxplot(outlier.color="brown",outlier.size=2
```

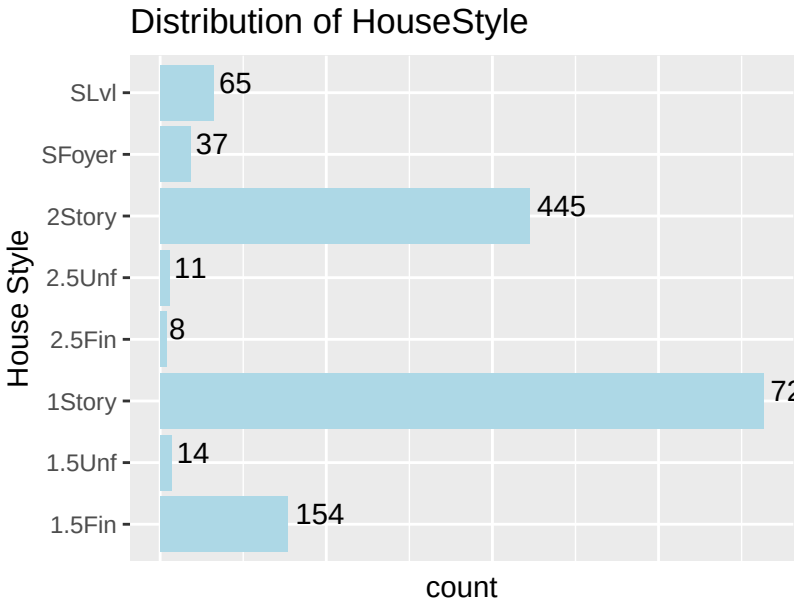
```
                )+scale_x_continuous(labels=label_number(scale=1e-3,suffix="k"))+
  theme(axis.text.y = element_blank(),
        axis.ticks.y=element_blank())
```



Description: Firstly we analyze a single variable(SalesPrice) which is also our main (primary) variable. So, here we got that minimum selling Price of houses which is 34.9k USD and maximum price of house is 75.5k USD whereas most of the houses are have selling price of 16.3k USD.

2. Analysis of Style of House

```
HouseSTYLE <- housing$HouseStyle
ggplot(housing,aes(y= HouseSTYLE))+labs(y="House Style")+
  ggtitle("Distribution of HouseStyle")+geom_bar(fill="lightblue")+
  geom_text(stat='count', aes(label=after_stat(count)), vjust=0,hjust=-0.15) +
  theme(axis.text.x = element_blank(),
        axis.ticks.x=element_blank())
```



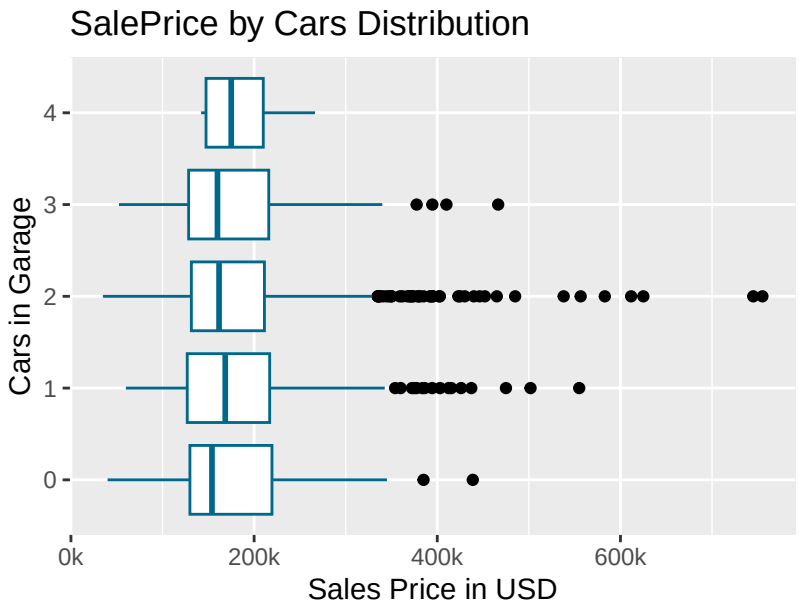
Description: Now we analyze another important single variable (i.e- HouseStyle) which represents style of house. In the above Bar-plot we can easily analyze that "1 Story" has maximum number of houses whereas the least is "2.5 Finished". Each bar depicts the count of each type of house style.

Till Now: Till now we discovered that the maximum and minimum price range of the houses and also count of House styles of housing data.

1.2 Advance Analysis between two variables

1. Analysis between Sale price of house and Number cars fits in a Garage.

```
PlotA <- ggplot(housing, aes(y = as.factor(GarageCars), x = SalePrice,
                             )) +
  geom_boxplot(outlier.color="black",color="deepskyblue4") +
  labs(x = "Sales Price in USD", y = "Cars in Garage") +
  ggtitle("SalePrice by Cars Distribution")+
  scale_x_continuous(labels=label_number(scale=1e-3,suffix=   "k"))
PlotA
```

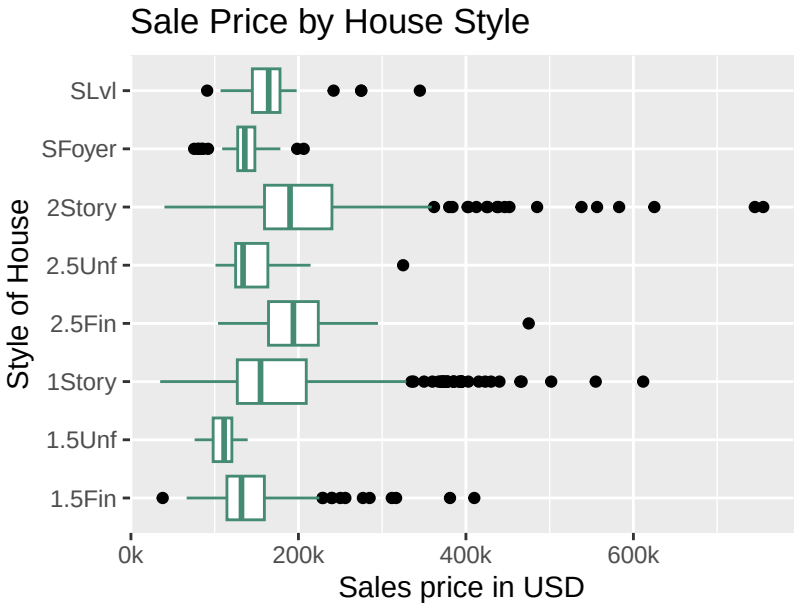


Description: Here we did Data visual analysis of two numerical variables (i.e - GarageCars and SalePrice). Where the first variable represents how many cars that can be parked in a garage and later one represents the Sale Price of that house respectively. Here as we can see the visual data, Each boxplot is grouped by the number of cars that can be fit in a garage. In terms of median, we have noticed that the house which has most number of cars space in garage has comparatively higher selling price than the ones which can fit less cars in their garage .There are some exceptions where garage that fits less number of cars has higher selling price, due to other factors that reflects and influences the total price of the house.Here, it can be seen that the houses with 2 Garage car space has the highest selling price. Garage with 0,1,2,3 number of cars have more than one outliers whereas the houses with 4 cars in its garage does not have any outliers.

2. Analysis between Sale price of house and Style of house.

```
Plot5 <- ggplot(housing, aes(x = SalePrice, y = HouseStyle)) +
  geom_boxplot(outlier.colour = "black",color="aquamarine4") +
  scale_x_continuous(labels=label_number(scale=1e-3,suffix=   "k")) +

  labs(x = "Sales price in USD",
  y = "Style of House") + ggtitle("Sale Price by House Style")
Plot5
```



Description: Here we did Data visual analysis of two numerical variables (i.e - Housestyle and SalePrice). The graph represent the sales price of houses on the basis of Style of house. Here as we can see the visual data in terms of median of boxplot and each boxplot represent a different HouseStyle. We can clearly see 2 Story House design has the maximum selling price while 1.5UNF(Unfurnished) are the most affordable ones. Each type of Houses have more than one outliers except 1.5UNF(Unfinished)

2 Conclusion

Each variable has a different attribute related to the house. We did the analysis between the Sales price of house with different variable affecting and influencing the price of the house. Firstly, we analysed two individual variable separately. Which shoes us some data, like number of each house style and maximum-minimum sales price of house followed by some advance analysis between our primary variable (i.e- Saleprice) with Secondary Variables (Garage Cars and HouseStyle). From these advance analysis, we got to know that the sales price of house is fully dependent upon these variables, and other secondary variables as well. To be more Precise, House with more space of cars in their garage having 2 Story as its housing style has the maximum sale price while on the other hand, the house with least space of car in garage and 1.5 unfinished as its housing style has the lower sale price.