Exploratory Data Analysis

Loading libraries

```
library(psych)
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.
3.0 --

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0

## -- Conflicts ------------------------------------------ tidyverse_conflict
s() --
## x ggplot2::%+%()   masks psych::%+%()
## x ggplot2::alpha() masks psych::alpha()
## x dplyr::filter()  masks stats::filter()
## x dplyr::lag()     masks stats::lag()

library(mice)

##
## Attaching package: 'mice'

## The following object is masked from 'package:stats':
##
##     filter

## The following objects are masked from 'package:base':
##
##     cbind, rbind

library(ggplot2)
library(knitr)
library(pastecs)

##
## Attaching package: 'pastecs'
```

```
## The following objects are masked from 'package:dplyr':
##
##     first, last

## The following object is masked from 'package:tidyr':
##
##     extract

library(dplyr)
```

Loading the data

```
housing.dataset<-read.csv("C:/Users/asus/Desktop/melbourne_data.csv")
```

1.Cleaning the Data

Removing columns with too many missing values

```
housing.dataset %>%
  select(-c(BuildingArea, YearBuilt)) -> housing.dataset
```

Replacing Landsize values that are zero with NA and removing outliers

```
housing.dataset$Landsize[housing.dataset$Landsize == 0] <- NA
housing.dataset$Landsize[housing.dataset$Landsize>30000] <- NA
```

Changing data types

```
housing.dataset$Date<-as.Date(housing.dataset$Date,"%d/%m/%y")
housing.dataset$Distance<-as.numeric(housing.dataset$Distance)

## Warning: NAs introduced by coercion

housing.dataset$Propertycount<-as.numeric(housing.dataset$Propertycount)

## Warning: NAs introduced by coercion
```

Removing all rows with either missing Price or Land size values

```
housing.dataset<-housing.dataset[!is.na(housing.dataset$Price)&!is.na(housing
.dataset$Landsize),]
```

Imputing missing values

```
tempData <- mice(housing.dataset, m=1, maxit=5, method='cart', seed=500)

##
##  iter imp variable
##   1   1  Bathroom  Car
##   2   1  Bathroom  Car
##   3   1  Bathroom  Car
##   4   1  Bathroom  Car
##   5   1  Bathroom  Car

## Warning: Number of logged events: 2
```

```r
housing.dataset <- complete(tempData,1)
```

2 Statistical analysis

Summary of Numerical Variables

```r
Desc.stat<-stat.desc(housing.dataset[,c(4:9,11)],basic=FALSE,desc =TRUE)
options(scipen=100)
options(digits=0)
knitr::kable(Desc.stat,caption="Table 1: Summary statistics of Numerical Vari
ables")
```

Table 1: Summary statistics of Numerical Variables

	Price	Landsize	Rooms	Bathroom	Car	Distance	Propertycount
median	970000	553	3	2	2	11	6482
mean	1150714	590	3	2	2	12	7383
SE.mean	5214	6	0	0	0	0	35
CI.mean.0.95	10221	12	0	0	0	0	68
var	435732457428	620185	1	1	1	45	19399961
std.dev	660100	788	1	1	1	7	4405
coef.var	1	1	0	0	1	1	1

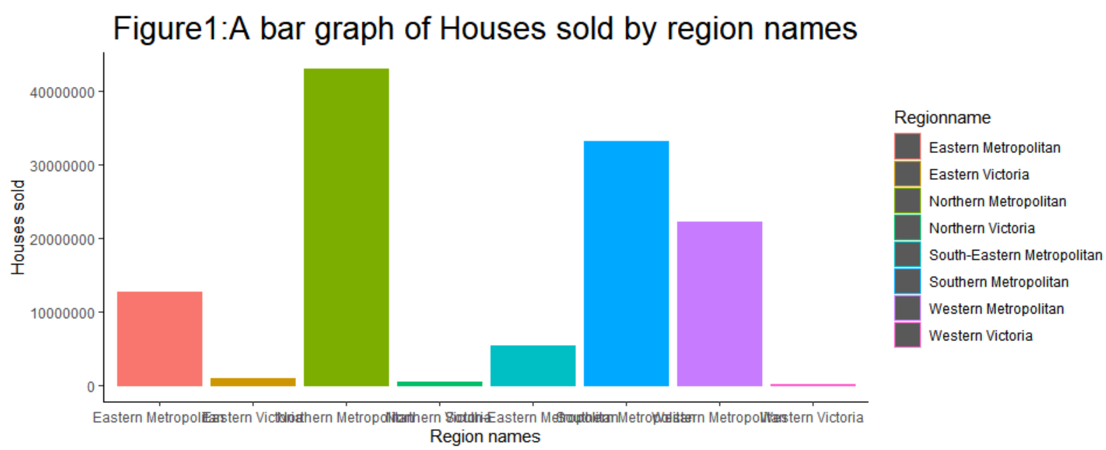
```r
# Dimensions of the cleaned data set.
dim(housing.dataset[,-1])

## [1] 16025    10

# nature of variables in the cleaned dataset
str(housing.dataset[,-1])

## 'data.frame':    16025 obs. of  10 variables:
##  $ Date        : Date, format: "2020-12-03" "2020-02-04" ...
##  $ Type        : chr  "h" "h" "h" "h" ...
##  $ Price       : int  1480000 1035000 1465000 850000 1600000 941000 18760
00 1636000 1097000 1350000 ...
##  $ Landsize    : int  202 156 134 94 120 181 245 256 220 214 ...
##  $ Rooms       : int  2 2 3 3 4 2 3 2 2 3 ...
##  $ Bathroom    : num  1 1 2 2 1 1 2 1 1 2 ...
##  $ Car         : num  1 0 0 1 2 0 0 2 2 2 ...
##  $ Distance    : num  2.5 2.5 2.5 2.5 2.5 2.5 2.5 2.5 2.5 2.5 ...
##  $ Regionname  : chr  "Northern Metropolitan" "Northern Metropolitan" "No
rthern Metropolitan" "Northern Metropolitan" ...
##  $ Propertycount: num  4019 4019 4019 4019 4019 ...
```

Our Cleaned dataset contains 10 variables and 16025 entries of Melbourne housing dataset. The data contained numerical, categorical, character and date types. The dataset contains the data collected from houses sold throughout the year 2020 across 8 districts in Melbourne, Australia. There is a sizable range between the 3rd quartile and the maximum value of most variable which leads us to believe our data is skewed by a few houses that vary from the others. The distribution is otherwise normal.

Variable analysis

Bar chart of Regions and number of houses sold there

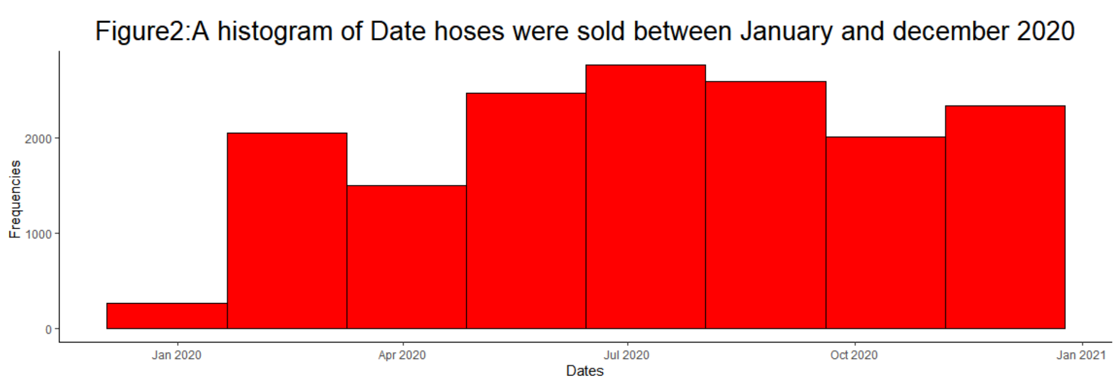
```
ggplot(housing.dataset,aes(y=Propertycount,x=Regionname,color=Regionname))+ge
om_bar(stat="identity")+theme_classic()+labs(title="Figure1:A bar graph of Ho
uses sold by region names")+theme(plot.title=element_text(hjust=0.5))+theme(p
lot.title=element_text(size=20))+ylab("Houses sold")+xlab("Region names")
```



##

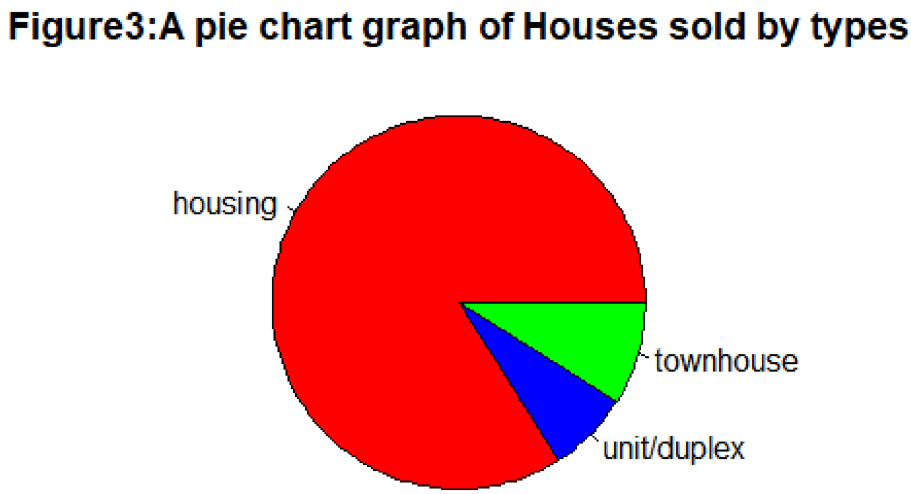
Distribution of houses sold in the reviewed time period - histogram

```
ggplot(housing.dataset,aes(x=Date))+geom_histogram(color="black",fill="red",b
ins=8)+theme_classic()+labs(title="Figure2:A histogram of Date hoses were sol
d between January and december 2020")+theme(plot.title=element_text(hjust=0.5
))+theme(plot.title=element_text(size=20))+ylab("Frequencies ")+xlab("Dates")
```



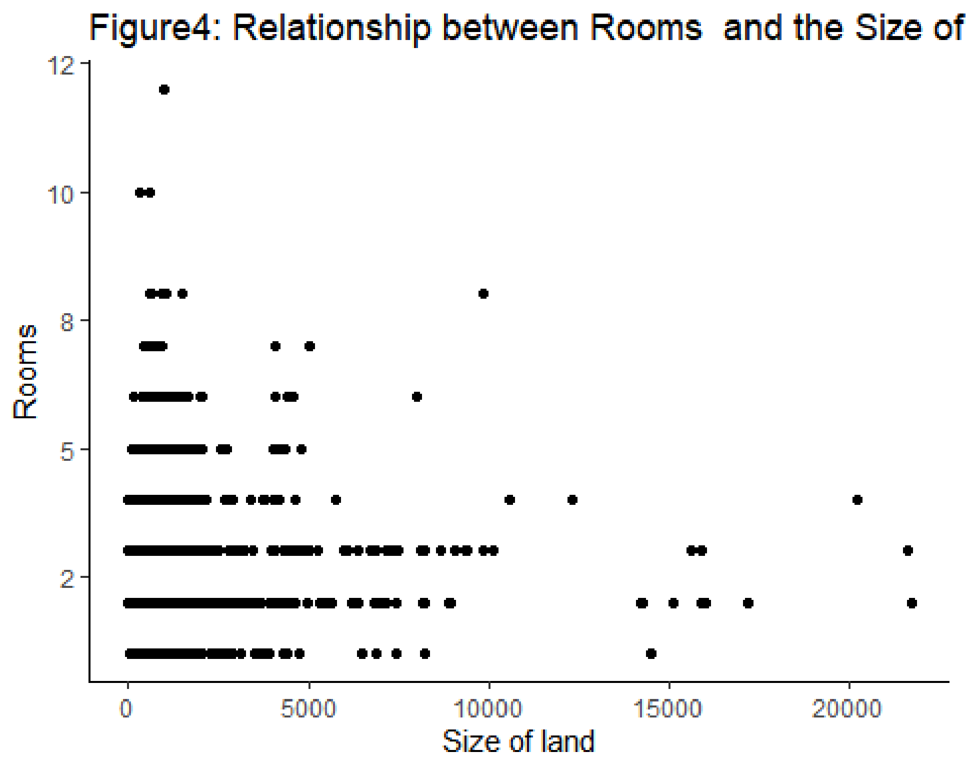
Piechart showing ratio of the different house types sold

```
#computing position of labels
pie(table(housing.dataset$Type), labels=c ("housing","unit/duplex","townhouse
"),radius=1,col=c("red","blue","green"),main="Figure3:A pie chart graph of Ho
uses sold by types")
```



Scatter plot of Landsize and number of rooms

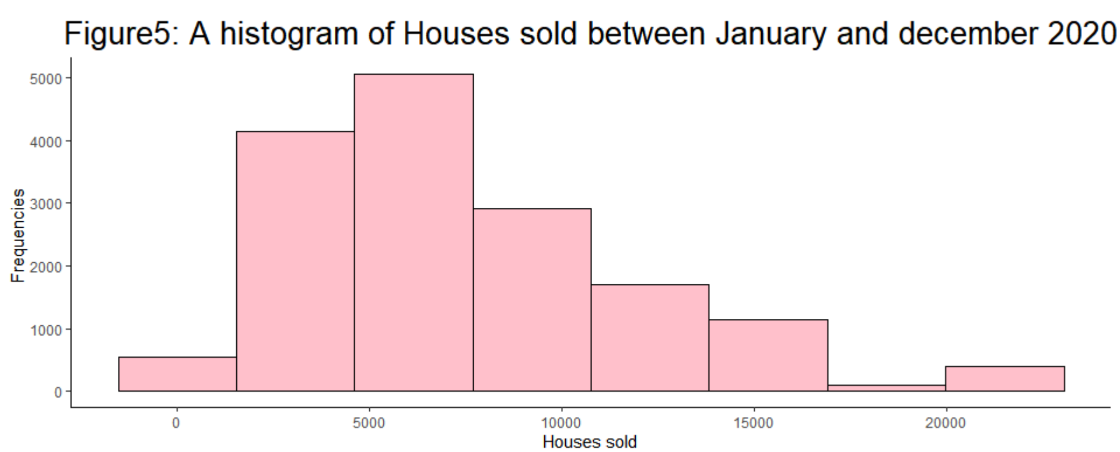
```
ggplot(housing.dataset, aes(x = Landsize, y =Rooms )) +
    geom_point()+theme_classic()+
    labs(
        x = "Size of land",
        y = "Rooms",

        title = "Figure4: Relationship between Rooms  and the Size of Land")
```



3. Analysis of Price Variable

Histogram of price variable

```
ggplot(housing.dataset,aes(x=Propertycount))+geom_histogram(color="black",fil
l="pink",bins=8)+theme_classic()+labs(title="Figure5: A histogram of Houses s
old between January and december 2020")+theme(plot.title=element_text(hjust=0
.5))+theme(plot.title=element_text(size=20))+ylab("Frequencies ")+xlab("House
s sold")
```



The histogram is heavily skewed to the left and indicates that a vast majority of the houses were sold for below 3 million Australian dollars.

Statistical analysis

```
Mean=mean(housing.dataset$Price,na.rm=TRUE)
SD=sd(housing.dataset$Price,na.rm=TRUE)
Median=median(housing.dataset$Price,na.rm=TRUE)
Variance<-var(housing.dataset$Price,na.rm=TRUE)
Lower_Quartile<-quantile(housing.dataset$Price,0.25)
Upper_Quartile<-quantile(housing.dataset$Price,0.75)
Skew<-skew(housing.dataset$Price)
kurtosis<-kurtosi(housing.dataset$Price)
summ<-data.frame(Mean,SD,Median,Variance,Lower_Quartile,Upper_Quartile,Skew,k
urtosis)
kable(summ,caption="Table 2: summary statistics of housing prices")
```

Table 2: summary statistics of housing prices

	Mean	SD	Median	Variance	Lower_Quartile	Upper_Quartile	Skew	kurtosis
25%	1150714	660100	970000	435732457428	711000	1400000	2	13

Group price by plot

```
summary(housing.dataset$Price)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##   131000   711000   970000  1150714  1400000 11200000
```

```
housing.dataset$Range<-ifelse(housing.dataset$Price<970000,"Low",ifelse(housi
ng.dataset$Price>=970000&housing.dataset$Price<=1400000,"Medium","High"))
```

Summary of price groups

```
# summary table of prices by Range
library(dplyr)
Summm=housing.dataset%>%
group_by(Range)%>%
  summarise(Obs=n(), Mean=mean(Price,na.rm=TRUE),SD=sd(Price,na.rm=TRUE))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

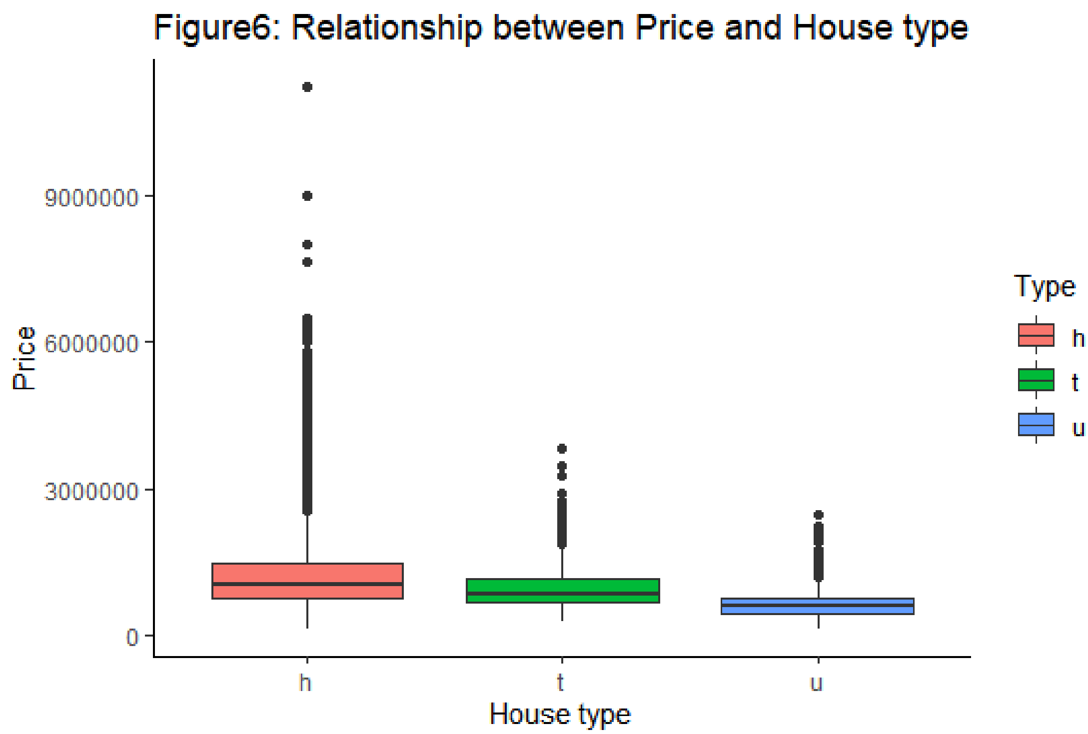
```
kable(Summm,caption="Table 3: summary statistics of housing prices Price Leve
ls")
```

Table 3: summary statistics of housing prices Price Levels

Range	Obs	Mean	SD
High	3948	2035399	712492
Low	7996	701839	161378
Medium	4081	1174350	128078

Exploring prices for different types of houses

```
housing.dataset %>%
  ggplot( aes(x=Type, y=Price, fill=Type)) +theme_classic()+
    geom_boxplot() +
    ggtitle("Figure6: Relationship between Price and House type") +
    xlab("House type")
```



##

Correletaion between price and other variables

```
i1<-sapply(housing.dataset,is.numeric)
y1<-"Price"
x1<-setdiff(names(housing.dataset)[i1],y1)
options(digits=5)
cor(housing.dataset[x1],housing.dataset[[y1]])
```

```
##                   [,1]
## X            -0.039279
## Landsize      0.029802
## Rooms         0.386985
## Bathroom      0.405620
## Car           0.147890
## Distance     -0.309284
## Propertycount -0.038635
```

The variables that correlate most strongly with Price include number of Bathrooms,number of rooms and Distance. Distance has a negative correlation with price meaning the closer the house is to the Central Business District the more expensive it tends to be.

4

Listing frequency of house types

```
table(housing.dataset$Type)

##
##     h     t     u
## 13441  1164  1420
```

13,441 Houses 1,164 Townhouses 1420 Units/Duplexes

Scatterplots

```
ggplot(housing.dataset, aes(x = Price, y = Landsize)) +
    geom_point(aes(color = factor(Type)))+theme_classic()  +
    labs(
        x = "Price of houses",
        y = "Size of Land",
        color = "Type of house",
        title = "Relation between Price and Size of Land")
```