B565-Data Mining
Homework #1

Due on Tuesday, Jan 31, 2023 08:00 p.m.

Dr. H. Kurban

Aniruddho Swapan Chatterjee

February 1, 2023

Problem 1

The following problems have to do with metrics. In each case, prove or disprove the distance is a metric ($\mathbb{R}$ is the set of reals, and $\|X\|$ is the size of a finite set X.)

Ans.

If the below three properties are satisfied for any given distance function then it can be proved that the distance is a metric.

1. Positivity
   (a) Greater than Zero
   $$d(x,x) \geq 0 \ \forall \ x,y \tag{1}$$
   (b) Equality
   $$d(x,y) = 0 \text{ when } x = y \tag{2}$$
2. Symmetry
   $$d(x,y) = d(y,x) \ \forall \ x,y \tag{3}$$
3. Triangle Property
   $$d(x,y) + d(y,z) \geq d(x,z) \ \forall \ x,y \tag{4}$$

(a) Let $X \subset \mathbb{R}^n$ for positive integer $n > 0$. Define a distance $d : X \times X \to \mathbb{R}_{\geq 0}$ as

$$d(x,y) = max\{|x_i - y_i|\}, \forall i, 1 \leq i \leq n.$$

Ans.

Considering equation 1 for the above function and substituting x we get max0
d(x,x)=0
Therefore equation 1 is satisfied
Considering equation 2 for the above function where x = y,we get d(x,y)=max0
d(x,y)=0
Therefore equation 2 is satisfied
Considering equation 3 for the above function.
Let x=5 and y=8
$d(x,y) = max|5-8|$
d(x,y)=3
$d(y,x) = max|8-5|$
d(y,x)=3
Therefore equation 3 is satisfied
Considering the equation 4 for the above function.
$|x-z|$ can be re-written as $|x-y+y-z|$
Therefore $|x-z| \leq |x-y| + |y-z|$
Taking max on both sides as per the function given above we get
$max|x-z| \leq max|x-y| + max|y-z|$ which is nothing but
$d(x,z) \leq d(x,y) + d(y,z)$
Therefore equation 4 is satisfied

Since all three properties are satisfied, the given function is a metric.

(b) Let $c : \mathbb{R}^{2n} \to \mathbb{R}_{\geq 0}$ be defined as

$$c(x,y) = \begin{cases} 1 & \text{if } x \neq y, \\ 0 & \text{o.w.} \end{cases}$$

Problem 1 continued on next page...

Define a distance $d : X \times X \to \mathbb{R}_{\geq 0}$ as

$$d(x, y) = \sum_{i}^{n} \frac{c(x_i, y_i)}{i}, \forall i \ 1 \leq i \leq n$$

Ans.

Considering equation 1 for the above function and substituting x for y we get c(x,x)=0 since x = y = x

Since c(x,y) = 0 , d(x,x)=0

Therefore equation 1 is satisfied

Considering equation 2 for the above function where x = y,we get c(x,y)=0

Since c(x,y) = 0 , d(x,y)=0

Therefore equation 2 is satisfied

Considering equation 3 for the above function.

When $x \neq y$ $c(x, y) = c(y, x) = 1$

$$d(x, y) = \sum_{i}^{n} \frac{1}{i}$$

$$d(y, x) = \sum_{i}^{n} \frac{1}{i}$$

Therefore equation 3 is satisfied

Considering the equation 4 for the above function.

When $x \neq y \neq z$ $c(x, y) = c(y, z) = c(x, z) = 1$

$$d(x, z) = \sum_{i}^{n} \frac{1}{i}$$

$$d(x, y) = \sum_{i}^{n} \frac{1}{i}$$

$$d(y, z) = \sum_{i}^{n} \frac{1}{i}$$

Clearly, adding two similar terms will be greater than the third.

Therefore equation 4 is satisfied

Since all three properties are satisfied, the given function is a metric.

(c) Suppose d_0, d_1 are metrics.

i. $d_0 \times d_1$

Ans.

Let us consider d_0 as $d_0(x, y) = max|x - y|$ and $d_1(x, y) = min(1, |y - x|)$

Using Equation 1 we get, $d_0(x, x) = 0$ *and* $d_1(x, x) = 0$

Therefore, $d_0 \times d_1 = 0$

Using Equation 2 if x=y we get, $d_0(x, y) = 0$ *and* $d_1(x, y) = 0$

Therefore, $d_0 \times d_1 = 0$

Therefore equation 2 is satisfied

Using Equation 3 let us consider x=8 and y =5,

$d_0(x, y) = 3$ *and* $d_1(x, y) = 1$

Therefore, $d_0 \times d_1 = 3$

In the same way, $d_0(y, x) = 3$ *and* $d_1(y, x) = 1$

Therefore equation 3 is satisfied

Using Equation 4 let us consider x=14 and y=15 and z=15.5,
$d_0(x,z) = 1.5$ $d_0(x,y) = 1 and$ $d_0(y,z) = 0.5$
$d_1(x,z) = 1$ $d_1(x,y) = 1 and$ $d_1(y,z) = 0.5$
Therefore, $d_0(x,z) \times d_1(x,z) = 1.5$
$d_0(x,y) \times d_1(x,y) + d_0(y,z) \times d_1(y,z) = 1.25$
Therefore equation 4 is not satisfied
Therefore $d_0 \times d_1$ is not a metric

ii. $(d_0 + d_1)/(d_0 d_1)$
Ans.
Using Equation 1 we get, $d_0(x,x) = 0$ *and* $d_1(x,x) = 0$
Therefore, $(d_0 + d_1)/(d_0 d_1) = \frac{0}{0}$ i.e undefined
Therefore $(d_0 + d_1)/(d_0 d_1)$ is not a metric

iii. $max\{d_0,\ d_1\}$
Ans.
Using Equation 1 we get, $d_0(x,x) = 0$ *and* $d_1(x,x) = 0$
Therefore, $max\{d_0,\ d_1\} = 0$

Using Equation 2 if x=y we get, $d_0(x,y) = 0$ *and* $d_1(x,y) = 0$
Therefore, $max\{d_0,\ d_1\} = 0$
Therefore equation 2 is satisfied

Using Equation 3 let us consider x=8 and y =5,
$d_0(x,y) = 3$ *and* $d_1(x,y) = 1$
Therefore, $max\{d_0,\ d_1\} = 3$
In the same way, $d_0(y,x) = 3$ *and* $d_1(y,x) = 1$
Therefore, $max\{d_0,\ d_1\} = 3$ for both the cases
Therefore equation 3 is satisfied

Using Equation 4 let us consider x=14 and y=15 and z=15.5,
$d_0(x,z) = 1.5$ $d_0(x,y) = 1 and$ $d_0(y,z) = 0.5$
$d_1(x,z) = 1$ $d_1(x,y) = 1 and$ $d_1(y,z) = 0.5$
Therefore, $max(d_0(x,z), d_1(x,z)) = 1.5$ and $max(d_0(x,y) + d_0(y,z), d_1(x,y) + d_1(y,z)) = 1.5$
Therefore equation 4 is satisfied
Therefore $max\{d_0,\ d_1\}$ is a metric

iv. Let X be a finite set. Define a distance $d : X \times X \to \mathbb{R}_{\geq 0}$ as $d(x,y) = \frac{||x \cap y||}{||x \cup y|| + 1}$
Ans.
Consider $x = 1, 2, 3, 4$ and $y = 5, 6$ Using Equation 1 we get, $d(x,x) = 0$
Using Equation 2 if x=y we get, $d(x,y) = 0$
Therefore equation 2 is satisfied

Using Equation 3
$d(x,y) = \frac{6}{7}$ *and* $d(y,x) = \frac{6}{7}$

Therefore equation 3 is satisfied

Using Equation 4 we get and consider $z = 1, 2, 3$
$d(x, z) = \frac{1}{8}$,$d(x, y) = 0$ and $d(y, z) = 0$
Therefore equation 4 is not satisfied

Therefore given equation is not a metric

Problem 2

Curse of Dimensionality: Generate m-dimensional n data points from a uniform distribution with values between 0 and 1. For an arbitrary m value

$$f(m) = \log_{10} \frac{d_{max}(m) - d_{min}(m)}{d_{min}(m)}$$

where $d_{max}(m)$ and $d_{min}(m)$ are maximum and minimum distances between any pair of points, respectively. Let m take each value from $\{1, 2, \ldots, 99, 100\}$. Repeat each experiment multiple times to get stable values by averaging the quantities over multiple runs for each m. For four different n values, e.g., $n \in \{150, 1500, 15000, 150000\}$, plot $f(m)$. Use Euclidean as your distance metric. Label and scale each axis properly and discuss your observations over different n's.

```
# Sample Python Script With Highlighting
import numpy as np
import matplotlib.pyplot as plt

def functionFM(minDist,maxDist):
  x = (maxDist - minDist)/(minDist)
  fm = np.log10(x)
  return fm

nValue=150
trails=10

def probabDist():
  dist=[]
  AvgFm=[]
  finalFm=[]
  for m in range(1,100):
    for t in range(1,trails):
      p=np.random.uniform(low = 0, high = 1, size = m)
      minDist=0
      maxDist=0
      dist=[]
      for n in range(1, nValue):
        l1=np.random.uniform(low = 0, high = 1, size = m)
        dist.append(np.linalg.norm(l1-p))
      minDist=min(dist)
      maxDist=max(dist)
      finalFm.append(functionFM(minDist,maxDist))
    AvgFm.append(np.average(finalFm))
```

```
  return AvgFm

plt.plot(np.arange(1,100).tolist(),probabDist(),label = "plot line")
plt.xlabel('m')
plt.ylabel('f(m)')
plt.title('Curse of dimensionality')
plt.legend()
plt.show()

nValue=1500
trails=10

plt.plot(np.arange(1,100).tolist(),probabDist(),label = "plot line")
plt.xlabel('m')
plt.ylabel('f(m)')
plt.title('Curse of dimensionality')
plt.legend()
plt.show()

nValue=15000
trails=10

plt.plot(np.arange(1,100).tolist(),probabDist(),label = "plot line")
plt.xlabel('m')
plt.ylabel('f(m)')
plt.title('Curse of dimensionality')
plt.legend()
plt.show()

nValue=150000
trails=10

plt.plot(np.arange(1,100).tolist(),probabDist(),label = "plot line")
plt.xlabel('m')
plt.ylabel('f(m)')
plt.title('Curse of dimensionality')
plt.legend()
plt.show()
```

Discussion of Experiments

Ans.

As the number of features in a data set increases the graph tends to approach zero quickly as it becomes steeper and steeper. This is the main reason of curse of dimensionality when the number of parameter values in the given space increases exponentially which also leads to consumption of time and resources for computation.

Plot/s

Place images here with suitable captions.

Figure 1: N=150 Trails = 10

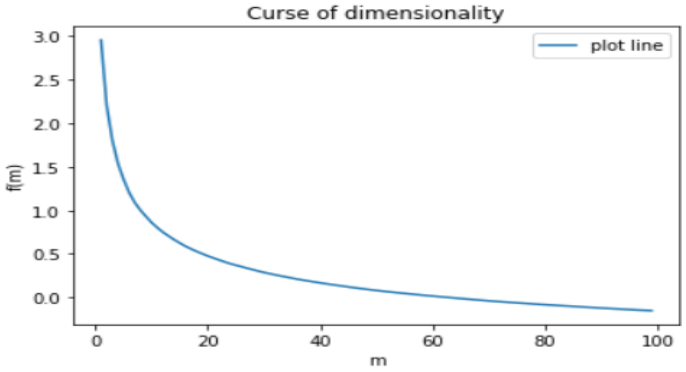


Figure 2: N=1500 Trails = 10

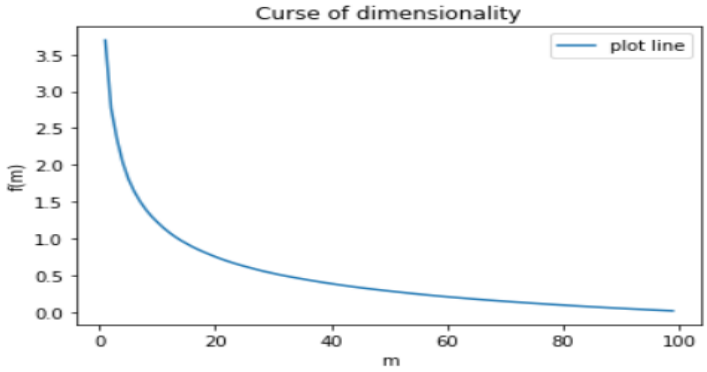


Figure 3: N=15000 Trails = 10

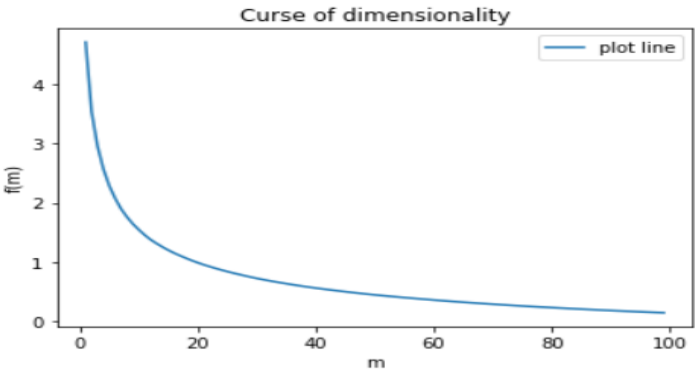
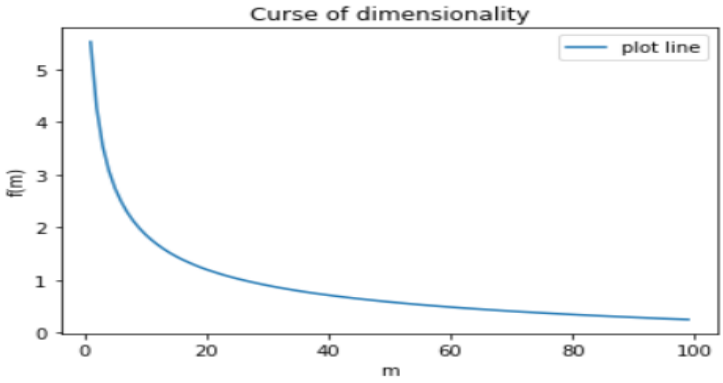


Figure 4: N=150000 Trails = 10



Problem 3

For the following data, give the best taxonomic type (interval, ratio, nominal, ordinal):

1. A section of highway on a map.
   Nominal
2. The value of a stock.
   Ratio
3. The weight of a person.
   Ratio
4. Marital status.
   Nominal
5. Visiting United Airlines (https://www.united.com) the seating is: Economony, Economy plus, and United Business.
   Ordinal

Problem 4

You are datamining with a column that has physical addresses in some city with the same zipcode. For example,

```
55 WEST CIR
2131 South Creek Road
Apt. #1 Fountain Park
1114 Rosewood Cir
1114 Rosewood Ct.
1114 Rosewood Drive
```

What structure would you create to mine these? What questions do you think you should be able to answer?
Ans.
We can structure this column into Apartment number, Street Number, Street Name, Zipcode, and city.
In this way we will be able to mine the data based on different attributes and gather useful information such as no of apartments on a particular street, the total number of apartments corresponding to a zipcode, and the total number of apartments in the city.
We can plot these on the graph as well and create better visualization to analyze the data effectively.

Problem 5

For this problem you will be using a data set with total 81 attributes describing every aspect of residential homes of Ames, Iowa. You can download the data from here [link]. The downloadable file already comes with the corresponding names of the attributes. Also a document describing the data is available here [link].

Discussion of Data

Briefly describe this data set–what is its purpose? How should it be used? What are the kinds of data it's using?
Ans.
The data set basically depicts the various features of a house in a particular area and how those features are affecting the Sales price of the house. The purpose is to predict sales price new house in that area based on the features.
The data is using both categorical as well as continuous variables/attributes such as OverallQuality Rating (1-10) and Number of bath (1-4), Living Area(0-6000 sq ft) and others.

R/Python Code

Using R/Python, show code that answers the following questions:

1. How many entries are in the data set? Write the R or Python code in the box below.
   Ans.
   There are 1460 records in the data set.

```python
# Sample Python Script With Highlighting
import csv
import pandas as pd
import matplotlib.pyplot as plt
from matplotlib.pyplot import figure
houseData=pd.read_csv('housing_data.csv')
print(len(houseData))
```

2. How many unknown or missing data are in the data set? Write the R or Python code in the box below.
   Ans.
   There are in total 6965 missing values.

```python
# Sample Python Script With Highlighting
#To get column wise number of null values
houseData.isnull().sum()
#To get total number of null values
houseData.isnull().sum().sum()
```

3. Find 10 attributes influencing the target attribute `SalePrice`. Use coherent plotting methods to describe and discuss their relation with `SalePrice`. Place images of these plots into the document. Write the R or Python code in the box below.

Ans.

Below are the 10 attributes influencing SalePrice.
OverallQual
GrLivArea
GarageCars
GarageArea
TotalBsmtSF
1stFlrSF
FullBath
TotRmsAbvGrd
YearBuilt
YearRemodAdd

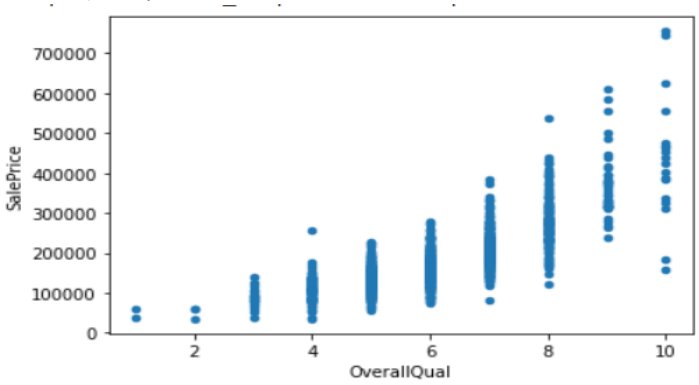
```
# Sample Python Script With Highlighting
unstackedCorr=houseData.corr().unstack().sort_values(ascending=False)
salesPriceCorr=unstackedCorr['SalePrice']
print(salesPriceCorr[0:11])#11 because Correlation of SalePrice to SalePrice is 1

#We can use Scatter plotting technique to show the relation
#between these 10 attributes and SalePrice

#Overall Quality Scatter Plot
houseData.plot(y='SalePrice', x='OverallQual', kind='scatter')
#Greater Living Area Scatter Plot
houseData.plot(y='SalePrice', x='GrLivArea', kind='scatter')
#Garage Cars Scatter Plot
houseData.plot( y='SalePrice', x='GarageCars', kind='scatter')
#Garage Area Scatter Plot#
houseData.plot(y='SalePrice', x='GarageArea', kind='scatter')
#Total Basement Square Feet Scatter Plot
houseData.plot(y='SalePrice', x='TotalBsmtSF', kind='scatter')
#First Floor Square Feet Scatter Plot
houseData.plot(y='SalePrice', x='1stFlrSF', kind='scatter')
#Full Bath Scatter Plot
houseData.plot( y='SalePrice', x='FullBath', kind='scatter')
#Total RMS Above Ground Scatter Plot
houseData.plot(y='SalePrice',x='TotRmsAbvGrd', kind='scatter')
#Year Built Scatter Plot
houseData.plot( y='SalePrice', x='YearBuilt', kind='scatter')
#Year Remod Add Scatter Plot
houseData.plot( y='SalePrice', x='YearRemodAdd', kind='scatter')
```

Discussion of Findings

Plot/s

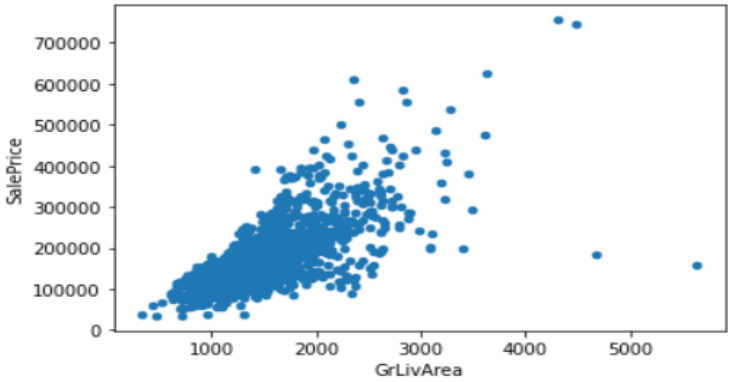
Place images here with suitable captions.
Ans.

(a) Overall Quality



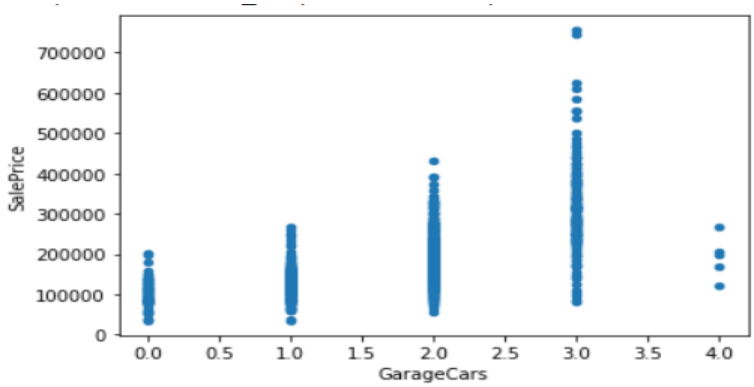
It is evident that Overall Quality is the most influential attribute for SalePrice with a correlation value of 0.79.
As and when the QverallQual value increases the price also increases.

(b) Greater Living Area



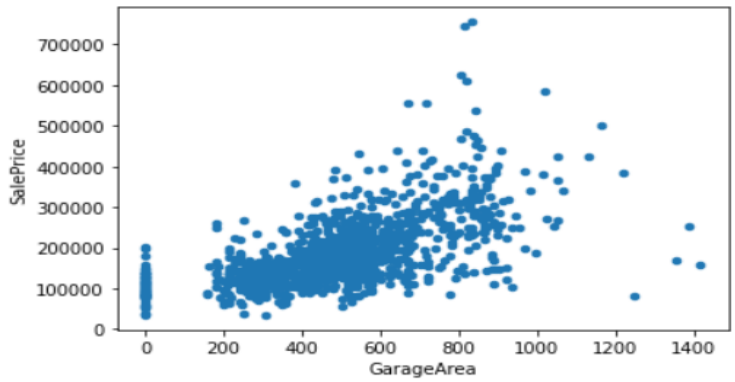
SalePrice is also increasing as the GrLivArea increases.
However, it is most dense in the range of 2000-3000 with SalePrice ranging from 100000 - 300000.

(c) Garage Cars



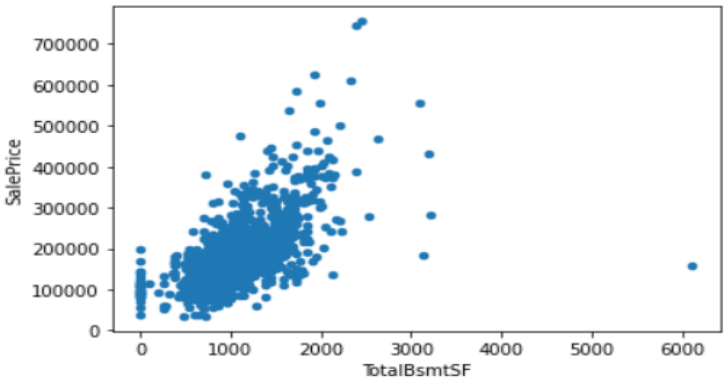
Houses below the range of 20000 do not have Garage Cars.
It can be inferred that there are very few houses that have Garage Cars above 3.

(d) Garage Area



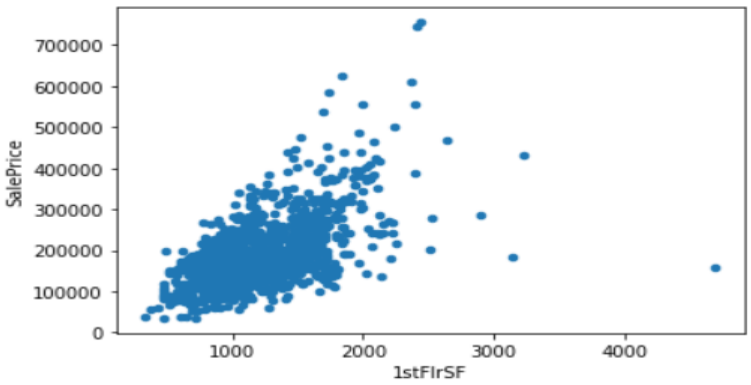
Also, Garage Area is not awarded to houses with SalesPrice less than 200000.

(e) Total Basement Square Feet



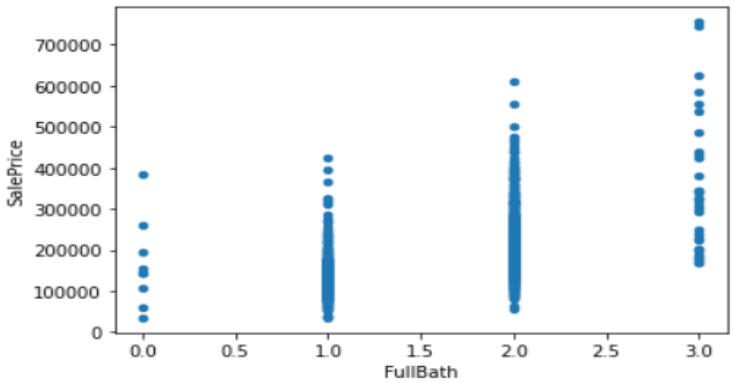
Also, TotalBsmtSF is not awarded to houses with SalesPrice less than 200000.

(f) First Floor Square Feet



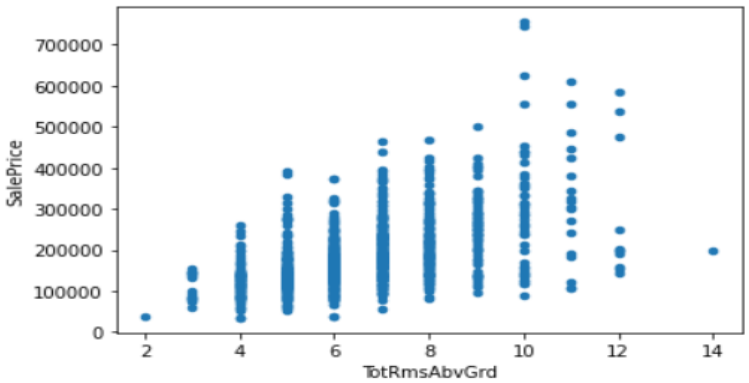
1stFlrSF increases as SalePrice increases.

(g) Full Bath



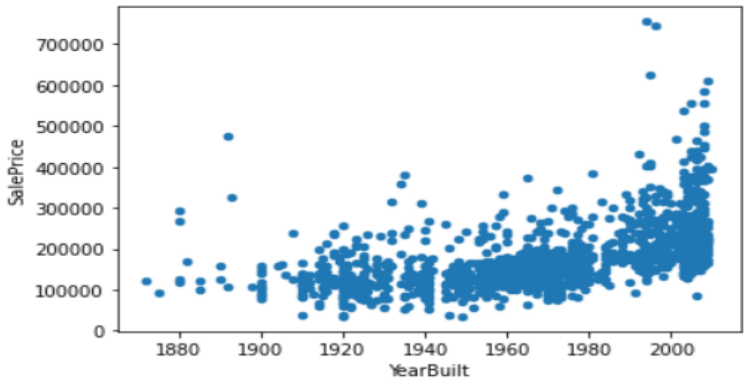
Most of the houses have 1-2 full baths with very few houses having 3 and others with no full bath.

(h) Total Rms Above Ground

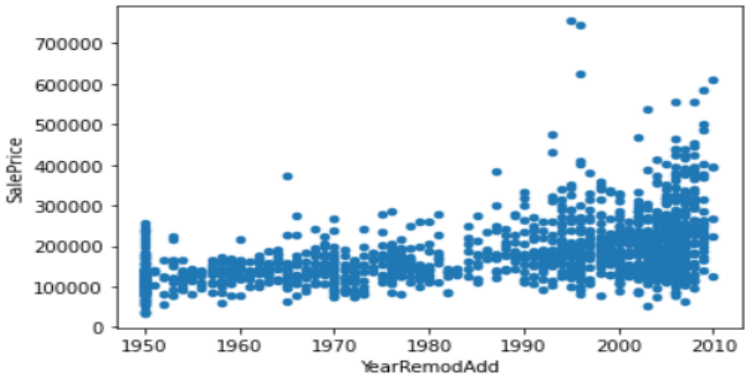


Total RMS Above Ground is somewhat uniform with some outliers where most of the houses have 4-10 with uniform distribution.

(i) Year Built



Houses with SalePrice more than 200000 were approximately built after 1990.



(j) YearRemodAdd

Houses that were built before 1940 were most likely Remodeled between 1950-1990.

4. Make a histogram/bar plot for each of those 10 attributes influencing `SalePrice` and discuss the distribution of values, *e.g.*, are uniform, skewed, normal of those attributes. Place images of these histograms into the document. Show the R/Python code that you used below and discussion below that.

```
# Sample Python Script With Highlighting
#Overall Quality Histogram

x1 = houseData.loc[houseData.OverallQual==1, 'SalePrice']
x2 = houseData.loc[houseData.OverallQual==2, 'SalePrice']
x3 = houseData.loc[houseData.OverallQual==3, 'SalePrice']
x4 = houseData.loc[houseData.OverallQual==4, 'SalePrice']
x5 = houseData.loc[houseData.OverallQual==5, 'SalePrice']
x6 = houseData.loc[houseData.OverallQual==6, 'SalePrice']
x7 = houseData.loc[houseData.OverallQual==7, 'SalePrice']
x8 = houseData.loc[houseData.OverallQual==8, 'SalePrice']
x9 = houseData.loc[houseData.OverallQual==9, 'SalePrice']
x10 = houseData.loc[houseData.OverallQual==10, 'SalePrice']

plt.hist(x1,alpha=0.5, bins=50,label='1')
plt.hist(x2,alpha=0.5, bins=50,label='2')
plt.hist(x3,alpha=0.5, bins=50,label='3')
plt.hist(x4,alpha=0.5, bins=50,label='4')
plt.hist(x5,alpha=0.5, bins=50,label='5')
```

```
plt.hist(x6,alpha=0.5, bins=50,label='6')
plt.hist(x7,alpha=0.5, bins=50,label='7')
plt.hist(x8,alpha=0.5, bins=50,label='8')
plt.hist(x9,alpha=0.5, bins=50,label='9')
plt.hist(x10,alpha=0.5, bins=50,label='10')
plt.gca().set(title='Frequency Histogram of Overall Quality'
,xlabel='Sales Price', ylabel='Frequency')
plt.legend();
plt.figure(figsize=(12,8))
plt.show();

#Greater Living Area Histogram

x1 = houseData.loc[houseData.GrLivArea.between(0,1000,inclusive='both')
, 'SalePrice']
x2 = houseData.loc[houseData.GrLivArea.between(1001,2000,inclusive='both')
, 'SalePrice']
x3 = houseData.loc[houseData.GrLivArea.between(2001,3000,inclusive='both')
, 'SalePrice']
x4 = houseData.loc[houseData.GrLivArea.between(3001,4000,inclusive='both')
, 'SalePrice']
x5 = houseData.loc[houseData.GrLivArea.between(4001,5000,inclusive='both')
, 'SalePrice']
x6 = houseData.loc[houseData.GrLivArea.between(5000,6000,inclusive='both')
, 'SalePrice']

plt.hist(x1,alpha=0.5, bins=50,label='0-1000')
plt.hist(x2,alpha=0.5, bins=50,label='1001-2000')
plt.hist(x3,alpha=0.5, bins=50,label='2001-3000')
plt.hist(x4,alpha=0.5, bins=50,label='3001-4000')
plt.hist(x5,alpha=0.5, bins=50,label='4001-5000')
plt.hist(x6,alpha=0.5, bins=50,label='5001-6000')
plt.gca().set(title='Frequency Histogram of Greater Living Area'
,xlabel='Sales Price', ylabel='Frequency')
plt.legend(title='Greater Living Area');
plt.show();

#Garage Cars Histogram

x1 = houseData.loc[houseData.GarageCars==1, 'SalePrice']
x2 = houseData.loc[houseData.GarageCars==2, 'SalePrice']
x3 = houseData.loc[houseData.GarageCars==3, 'SalePrice']
x4 = houseData.loc[houseData.GarageCars==4, 'SalePrice']

plt.hist(x1,alpha=0.5, bins=50, color='r', label='1')
plt.hist(x2,alpha=0.5, bins=50, color='g', label='2')
plt.hist(x3,alpha=0.5, bins=50, color='b', label='3')
plt.hist(x4,alpha=0.5, bins=50, color='magenta', label='4')
plt.gca().set(title='Frequency Histogram of Garage Cars'
,xlabel='Sales Price', ylabel='Frequency')
plt.legend(title='Garage Cars');
plt.show();
```

```
#Garage Area Histogram

x1 = houseData.loc[houseData.GarageArea.between(0,400,inclusive='both')
, 'SalePrice']
x2 = houseData.loc[houseData.GarageArea.between(401,800,inclusive='both')
, 'SalePrice']
x3 = houseData.loc[houseData.GarageArea.between(801,1200,inclusive='both')
, 'SalePrice']
x4 = houseData.loc[houseData.GarageArea.between(1201,1600,inclusive='both')
, 'SalePrice']

plt.hist(x1,alpha=0.5, bins=50,label='0-400')
plt.hist(x2,alpha=0.5, bins=50,label='401-800')
plt.hist(x3,alpha=0.5, bins=50,label='801-1200')
plt.hist(x4,alpha=0.5, bins=50,label='1201-1600')
plt.gca().set(title='Frequency Histogram of Garage Area'
,xlabel='Sales Price', ylabel='Frequency')
plt.legend(title='Garage Area');
plt.show();

#Total Basement Square Feet Histogram

x1 = houseData.loc[houseData.TotalBsmtSF.between(0,1000,inclusive='both')
, 'SalePrice']
x2 = houseData.loc[houseData.TotalBsmtSF.between(1001,2000,inclusive='both')
, 'SalePrice']
x3 = houseData.loc[houseData.TotalBsmtSF.between(2001,3000,inclusive='both')
, 'SalePrice']
x4 = houseData.loc[houseData.TotalBsmtSF.between(3001,4000,inclusive='both')
, 'SalePrice']
x5 = houseData.loc[houseData.TotalBsmtSF.between(4001,7000,inclusive='both')
, 'SalePrice']
plt.hist(x1,alpha=0.5, bins=50,label='0-1000')
plt.hist(x2,alpha=0.5, bins=50,label='1001-2000')
plt.hist(x3,alpha=0.5, bins=50,label='2001-3000')
plt.hist(x4,alpha=0.5, bins=50,label='3001-4000')
plt.hist(x4,alpha=0.5, bins=50,label='4001-7000')
plt.gca().set(title='Frequency Histogram of Total Basement Square Feet'
,xlabel='Sales Price', ylabel='Frequency')
plt.legend(title='Total Basement Square Feet');
plt.show();

#1st Floor Square Feet Histogram

x1 = houseData.loc[houseData['1stFlrSF'].between(0,1000,inclusive='both')
, 'SalePrice']
x2 = houseData.loc[houseData['1stFlrSF'].between(1001,2000,inclusive='both')
, 'SalePrice']
x3 = houseData.loc[houseData['1stFlrSF'].between(2001,3000,inclusive='both')
, 'SalePrice']
x4 = houseData.loc[houseData['1stFlrSF'].between(3001,4000,inclusive='both')
, 'SalePrice']
x5 = houseData.loc[houseData['1stFlrSF'].between(4001,7000,inclusive='both')
```

```
, 'SalePrice']
plt.hist(x1,alpha=0.5, bins=50,label='0-1000')
plt.hist(x2,alpha=0.5, bins=50,label='1001-2000')
plt.hist(x3,alpha=0.5, bins=50,label='2001-3000')
plt.hist(x4,alpha=0.5, bins=50,label='3001-4000')
plt.hist(x4,alpha=0.5, bins=50,label='4001-7000')
plt.gca().set(title='Frequency Histogram of Total First Floor Square Feet'
,xlabel='Sales Price', ylabel='Frequency')
plt.legend(title='Total First Floor Square Feet');
plt.show();

#Full bath Histogram

x1 = houseData.loc[houseData.FullBath==1, 'SalePrice']
x2 = houseData.loc[houseData.FullBath==2, 'SalePrice']
x3 = houseData.loc[houseData.FullBath==3, 'SalePrice']
x4 = houseData.loc[houseData.FullBath==4, 'SalePrice']

plt.hist(x1,alpha=0.5, bins=50, color='r', label='1')
plt.hist(x2,alpha=0.5, bins=50, color='g', label='2')
plt.hist(x3,alpha=0.5, bins=50, color='b', label='3')
plt.hist(x4,alpha=0.5, bins=50, color='magenta', label='4')
plt.gca().set(title='Frequency Histogram of Full Bath'
,xlabel='Sales Price', ylabel='Frequency')
plt.legend(title='Full Bath');
plt.show();

#Total RMS Above Ground Histogram

x1 = houseData.loc[houseData.TotRmsAbvGrd.between(0,4,inclusive='both')
, 'SalePrice']
x2 = houseData.loc[houseData.TotRmsAbvGrd.between(5,8,inclusive='both')
, 'SalePrice']
x3 = houseData.loc[houseData.TotRmsAbvGrd.between(9,14,inclusive='both')
, 'SalePrice']
x4 = houseData.loc[houseData.TotRmsAbvGrd.between(14,18,inclusive='both')
, 'SalePrice']

plt.hist(x1,alpha=0.5, bins=50,label='0-4')
plt.hist(x2,alpha=0.5, bins=50,label='5-8')
plt.hist(x3,alpha=0.5, bins=50,label='9-14')
plt.hist(x4,alpha=0.5, bins=50,label='14-18')
plt.gca().set(title='Frequency Histogram of Total Rms Above Ground'
,xlabel='Sales Price', ylabel='Frequency')
plt.legend(title='Total Rms Above Ground');
plt.show();

#Year Built Histogram

x1 = houseData.loc[houseData.YearBuilt.between(1800,1850,inclusive='both')
, 'SalePrice']
x2 = houseData.loc[houseData.YearBuilt.between(1851,1900,inclusive='both')
, 'SalePrice']
```

```
x3 = houseData.loc[houseData.YearBuilt.between(1901,1950,inclusive='both')
, 'SalePrice']
x4 = houseData.loc[houseData.YearBuilt.between(1951,2000,inclusive='both')
, 'SalePrice']

plt.hist(x1,alpha=0.5, bins=50,label='1800-1850')
plt.hist(x2,alpha=0.5, bins=50,label='1851-1900')
plt.hist(x3,alpha=0.5, bins=50,label='1901-1950')
plt.hist(x4,alpha=0.5, bins=50,label='1951-2000')
plt.gca().set(title='Frequency Histogram of Year Built'
,xlabel='Sales Price', ylabel='Frequency')
plt.legend(title='Year Built');
plt.show();

#Year Remod Add Histogram

x1 = houseData.loc[houseData.YearRemodAdd.between(1900,1950,inclusive='both')
, 'SalePrice']
x2 = houseData.loc[houseData.YearRemodAdd.between(1951,2000,inclusive='both')
, 'SalePrice']
x3 = houseData.loc[houseData.YearRemodAdd.between(2001,2050,inclusive='both')
, 'SalePrice']

plt.hist(x1,alpha=0.5, bins=50,label='1900-1950')
plt.hist(x2,alpha=0.5, bins=50,label='1951-2000')
plt.hist(x3,alpha=0.5, bins=50,label='2001-2050')
plt.gca().set(title='Frequency Histogram of Year Remod Add'
,xlabel='Sales Price', ylabel='Frequency')
plt.legend(title='Year Remod Add');
plt.show();
```

Discussion of Attributes

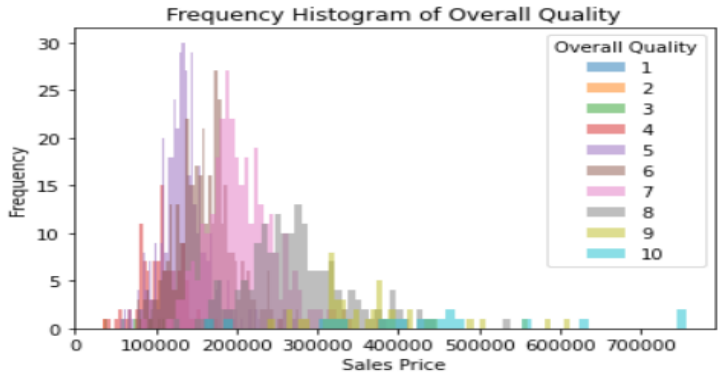
Histograms/Bar Plots

Place images here with suitable captions.

Ans.

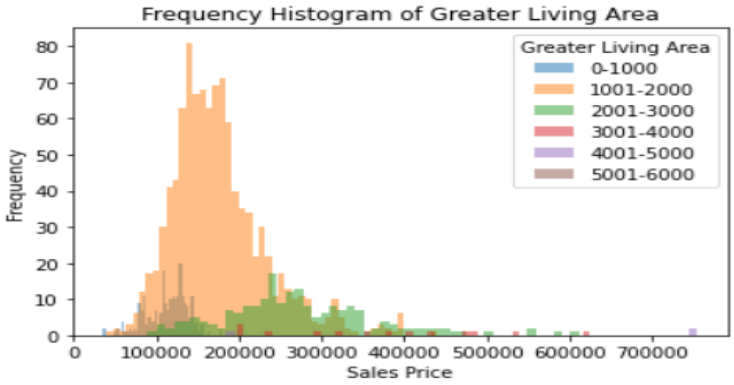
From the below histograms, it is evident that the housing data is mostly positively skewed for the top 10 attributes that are influencing SalePrice.

(a) Overall Quality



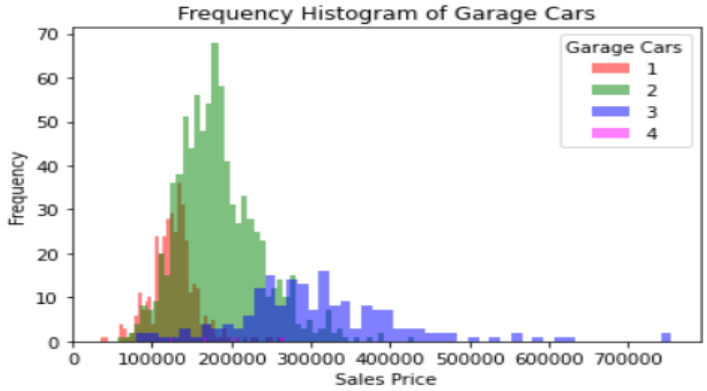
Majority of the frequencies of Overall Quality lie in the range of SalePrice of 100000-300000.

(b) Greater Living Area



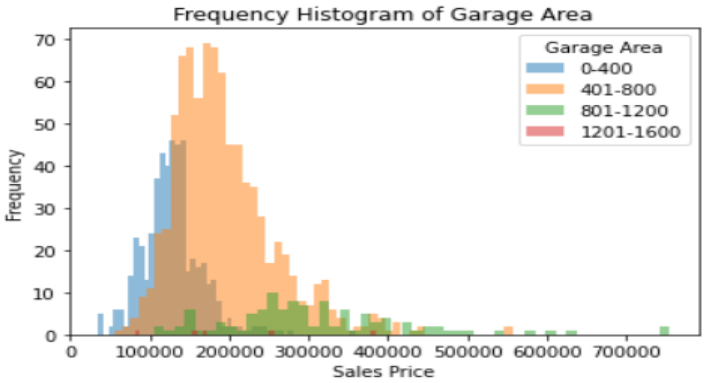
Sale Prices of houses in the range 100000-200000 have majority GrLivArea in the range of 1001-2000

(c) Garage Cars



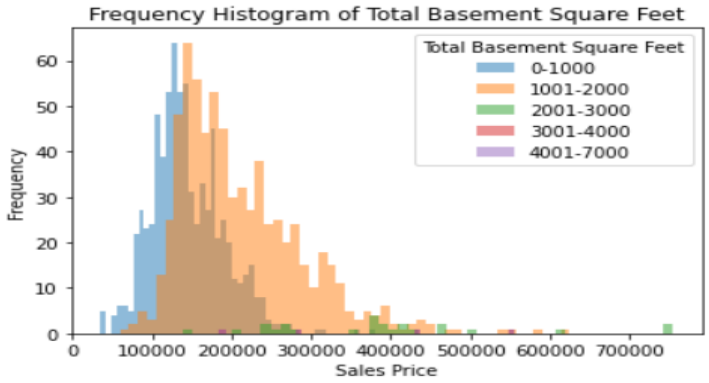
Majority of the houses have 2 Garage Cars followed by 1 Garage Car and then 3 Garage Cars Also, It can be inferred that there are very few houses that have Garage Cars above 3.

(d) Garage Area



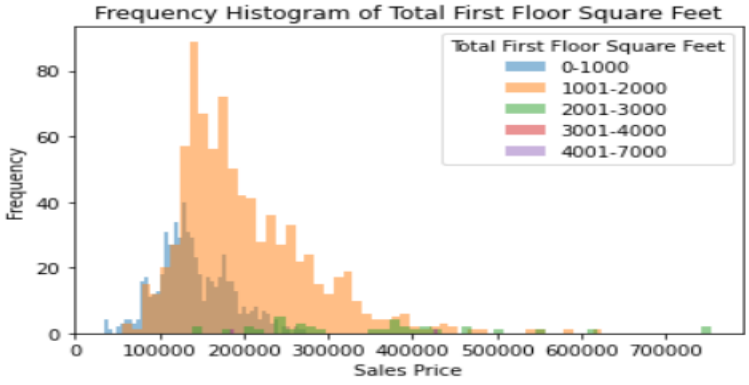
Majority of the houses have a Garage Area of 401-800 and Sale prices in the range of 100000-300000.

(e) Total Basement Square Feet



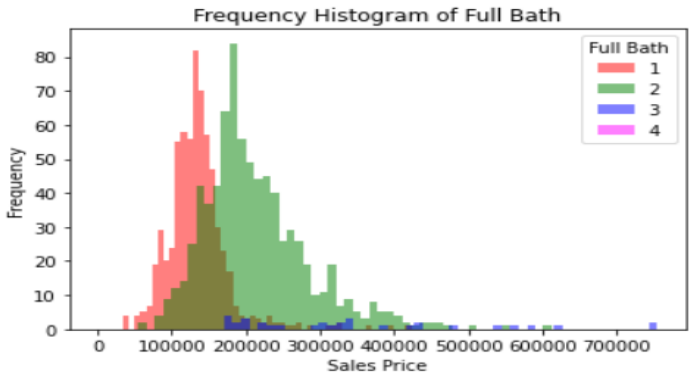
Majority of Total Basement Square Feet are in the range of 1001-2000 and the Sale price is in the range of 150000-250000.

(f) First Floor Square Feet



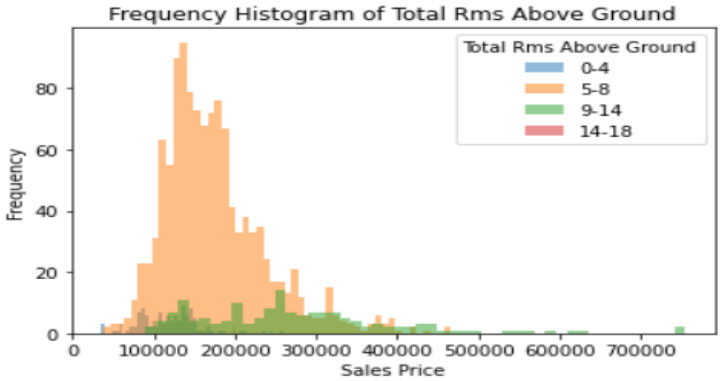
Majority of the houses have First Floor Square Feet in the range of 1001-2000.

(g) Full bath



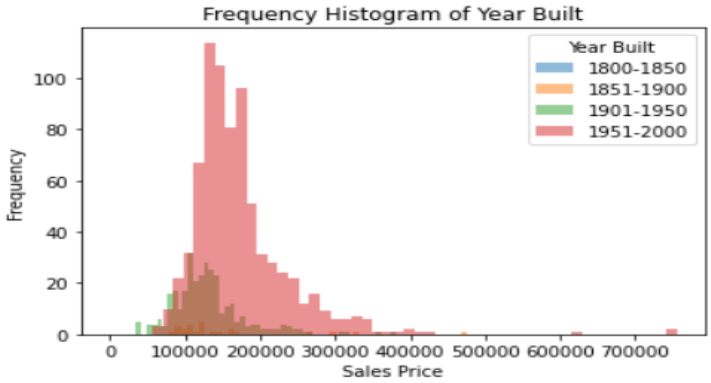
Most of the houses have 1-2 full baths with very few houses having 3 and others with no full baths.

(h) Total Rms Above Ground



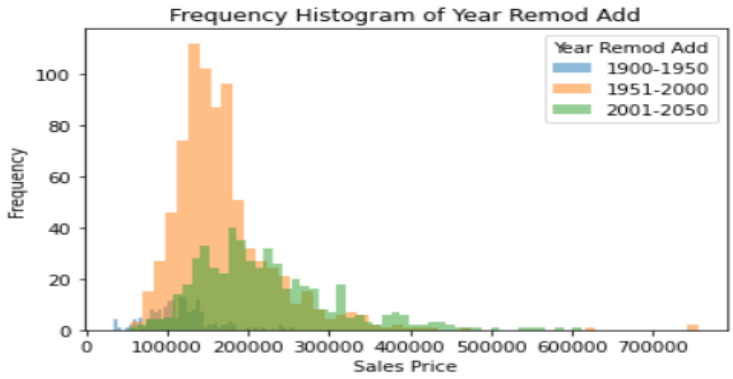
Majority of the houses have Total Rms Above Ground in the range 5-8.

(i) Year Built



Majority of the houses have been built in the range 1951-2000

(j) YearRemodAdd



Majority of the Houses were Remodeled between 1951-2000.

Discussion of simply removing tuples

Quantify the affect of simply removing the tuples with unknown or missing values. What is the cost in human capital?

Ans.

Simply removing the tuple with unknown or missing value might affect the other attributes which were being used to extract value from the data set.

We need to analyse first if the data is missing at random, if all the values of that column are missing or the data is corrupted.

Replacing with zero, mean or median would be a good alternative.

To begin with, one needs to understand the data at hand and have the skill sets to first import the data set and analyse. This will consume a lot of time and resources.

Problem 6

Distinguish between noise and outliers. Be sure to consider the following questions.

1. Is noise ever interesting or desirable? Outliers?
   Ans.
   Noise can be defined as distortions/corruption in the original data. Hence, noise is undesirable while working with data and should be handled in the pre-processing phase. It is also considered to be the primary reason for a model overfitting. Outliers, on the other hand, can be of some interest based on the goal one is trying to achieve.

2. Can noise objects be outliers?
   Ans.
   Yes, as mentioned earlier, noise is the distortion/corruption of data and this can happen anywhere in the data.

3. Are noise objects always outliers?
   Ans.
   No, distortions in data can also occur as normal data and not always necessarily as outliers.

4. Are outliers always noise objects?
   Ans.

No, outliers can be legitimate data without any distortions just that its representation is not as per the other values in the dataset.

5. Can noise make a typical value into an unusual one, or vice versa?
   Ans.
   Yes, introduction or presence of noise can make an impact on quality of data and make it unusual.

Problem 7

You are given a set of m objects that is divided into K groups, where the i^{th} group is of size m_i. If the goal is to obtain a sample of size $n < m$, what is the difference between the following sampling schemes? (Assume sampling with replacement.)

1. We randomly select $n * m_i/m$ elements from each group.
2. We randomly select n elements from the data set, without regard for the group to which an object belongs.

Ans.
Selecting randomly from each group refers to as Stratified Sampling technique. On the other hand, selecting random values from the data set refers to as Random Sampling technique.
Random sampling is used to get a generalised idea and analysis of the data set at hand. However, Stratified sampling gives more accurate results since it is first grouped based on some criteria and then selected for further analysis.
More resources are required to mine the whole data set in random sampling as compared to stratified sampling where analysis can be done on the small samples generated.
Random sampling is frequently used on data sets where there is no/less meaningful information(attributes) available and Stratified is used when the attributes are related to each other and information can be extracted.

Problem 8

Consider a document-term matrix, where tf_{ij} is the frequency of the i^{th} word (term) in the j^{th} document and m is the number of documents. Consider the variable transformation that is defined by

$$tf'_{ij} = tf_{ij} * \log \frac{m}{df_i},$$

where df_i is the number of documents in which the i^{th} term appears, which is known as the document frequency of the term. This transformation is known as inverse document frequency transformation.

1. What is the effect of this transformation if a term occurs in one document? In every document?
   Ans.
   If a particular term occurs in one document $df_i = 1$ then the transformation yields $\log m$ times.
   On the other hand, if the term occurs in every document then df_i = m and the transformation yields 0
2. What might be the purpose of this transformation?
   Ans.
   This kind of transformation is mostly used to identify the terms that are occurring rarely.
   The values of this transformation range from 0 to $\log m$.

Problem 9

This question compares and contrasts some similarity and distance measures.

1. For binary data, the L1 distance corresponds to the Hamming distance; that is, the number of bits that are different between two binary vectors. The Jaccard similarity is a measure of the similarity between two binary vectors. Compute the Hamming distance and the Jaccard similarity between the following two binary vectors.

   - $\mathbf{x} = 0101010001$
   - $\mathbf{y} = 0100011000$

   Ans.
   Hamming Distance = no of bits which is not the same.
   Hamming Distance = 3
   Jaccard Similarity can be given by the following equation

   $$Jaccard(x, y) = \frac{f_{11}}{f_{01} + f_{10} + f_{11}} \tag{5}$$

   $Jaccard(x, y) = \frac{2}{5} = 0.4$
   Jaccard(x,y) = 0.4

2. Which approach, Jaccard or Hamming distance, is more similar to the Simple Matching Coefficient, and which approach is more similar to the cosine measure? Explain. (Note: The Hamming distance is a distance, while the other three measures are similarities, but don't let this confuse you.)
   Ans.
   Jaccard Similarity is more similar to Cosine similarity co-efficient since we neglect the f_{00} in both of them.
   Hamming distance is more similar to Simple Matching Coefficient since we consider the f_{00} in both of them.

3. Suppose that you are comparing how similar two organisms of different species are in terms of the number of genes they share. Describe which measure, Hamming or Jaccard, you think would be more appropriate for comparing the genetic makeup of two organisms. Explain. (Note: Assume that each animal is represented as a binary vector, where each attribute is 1 if a particular gene is present in the organism and 0 otherwise.)
   Ans.
   In order to compare the genes of two organisms it will be better to use Jaccard, according to the equation 5, since we are looking for as many attributes where the gene is present and represented by f_{11}.
   Hamming distance however focuses more on the f_{00} frequency which corresponds to the genes which are absent.
   Hence, Jaccard approach would be better to determine presence of genes

4. If you wanted to compare the genetic makeup of two organisms of the same species, e.g., two human beings, would you use the Hamming distance, the Jaccard coefficient, or a different measure of similarity or distance? Explain. (Note: Two human beings share $> 99.9\%$ of the same genes.)
   Ans.
   If it is given that two human beings share $> 99.9\%$ of the same genes, then it would be better to use Hamming distance as a measure to compare the genetic makeup since Hamming will look for those genes which are not the same i.e $< 0.1\%$.
   Hence, our data set will be reduced and we will be able to work on it effectively.

Problem 10

For the following vectors, $\mathbf{x}$ and $\mathbf{y}$, calculate the indicated similarity and distance measures. Show detailed calculations/steps.

1. $\mathbf{x} = (1, 1, 1, 1)$, $\mathbf{y} = (2, 2, 2, 2)$ cosine, correlation, Euclidean.
2. $\mathbf{x} = (0, 1, 0, 1)$, $\mathbf{y} = (1, 0, 1, 0)$ cosine, correlation, Euclidean, Jaccard.
3. $\mathbf{x} = (0, -1, 0, 1)$, $\mathbf{y} = (1, 0, -1, 0)$ cosine, correlation, Euclidean.
4. $\mathbf{x} = (1, 1, 0, 1, 0, 1)$, $\mathbf{y} = (1, 1, 1, 0, 0, 1)$ cosine, correlation, Jaccard.
5. $\mathbf{x} = (2, -1, 0, 2, 0, -3)$, $\mathbf{y} = (-1, 1, -1, 0, 0 - 1)$ cosine, correlation.

Ans.

1. $\mathbf{x} = (1, 1, 1, 1)$, $\mathbf{y} = (2, 2, 2, 2)$.

   (a) Cosine
   Cosine similarity can be calculated using the below equation.

$$cosine(x, y) = \frac{x.y}{||x||.||y||} \tag{6}$$

   $x.y = 2 \times 1 + 2 \times 1 + 2 \times 1 + 2 \times 1 = 8$
   $||x|| = \sqrt{(1 \times 1) + (1 \times 1) + (1 \times 1) + (1 \times 1)} = 2$
   $||y|| = \sqrt{(2 \times 2) + (2 \times 2) + (2 \times 2) + (2 \times 2)} = 4$
   Therefore, cosine similarity is 1

   (b) Correlation.
   Correlation can be calculated using the below equation.

$$Corr(x, y) = \frac{Cov(x, y)}{\sigma_x.\sigma_y} \tag{7}$$

$$Cov(x, y) = \frac{\sum(x - \overline{x})(y - \overline{y})}{N} \tag{8}$$

$$\sigma_x = \sqrt{\frac{\sum(x - \overline{x})^2}{N}} \tag{9}$$

$$\sigma_y = \sqrt{\frac{\sum(y - \overline{y})^2}{N}} \tag{10}$$

   where ,
   $\overline{x}-$ Mean of x
   $\overline{y}-$ Mean of y
   In our case,
   $\overline{x} = 1$
   $\overline{y} = 2$
   $Cov(x, y) = \frac{(1-1)(2-2)+(1-1)(2-2)+(1-1)(2-2)+(1-1)(2-2)}{4} = 0$
   $\sigma_x = \sqrt{\frac{(1-1)^2+(1-1)^2+(1-1)^2+(1-1)^2}{4}} = 0$
   $\sigma_y = \sqrt{\frac{(2-2)^2+(2-2)^2+(2-2)^2+(2-2)^2}{4}} = 0$
   Therefore, correlation is $\frac{0}{0}$ i.e undefined

(c) Euclidean.

$$Euclidean(x,y) = \sqrt{\sum (y-x)^2} \tag{11}$$

$Euclidean(x,y) = \sqrt{(2-1)^2+(2-1)^2+(2-1)^2+(2-1)^2} = 2$
Therefore, Euclidean distance is 2

2. $\mathbf{x} = (0,1,0,1)$, $\mathbf{y} = (1,0,1,0)$.

(a) Cosine
Using equation 6
$x.y = 1\times 0+1\times 0+1\times 0+1\times 0 = 0$
$||x|| = \sqrt{(0\times 0)+(1\times 1)+(0\times 0)+(1\times 1)} = \sqrt{2}$
$||y|| = \sqrt{(1\times 1)+(0\times 0)+(1\times 1)+(0\times 0)} = \sqrt{2}$
Therefore, cosine similarity is 0

(b) Correlation.
Using equation 7,8,9,10
$\overline{x} = 0.5$
$\overline{y} = 0.5$
$Cov(x,y) = \frac{(0-0.5)(1-0.5)+(1-0.5)(0-0.5)+(0-0.5)(1-0.5)+(1-0.5)(0-0.5)}{4} = -0.25$
$\sigma_x = \sqrt{\frac{(0-0.5)^2+(1-0.5)^2+(0-0.5)^2+(1-0.5)^2}{4}} = 0.5$
$\sigma_y = \sqrt{\frac{(1-0.5)^2+(0-0.5)^2+(1-0.5)^2+(0-0.5)^2}{4}} = 0.5$
Therefore, correlation is $\frac{-0.25}{0.5\times 0.5}$ = -1

(c) Euclidean.
Using equation 11
$Euclidean(x,y) = \sqrt{(1-0)^2+(1-0)^2+(1-0)^2+(1-0)^2} = 2$
Therefore, Euclidean distance is 2

(d) Jaccard.

$$Jaccard(x,y) = \frac{f_{11}}{f_{01}+f_{10}+f_{11}} \tag{12}$$

$Jaccard(x,y) = \frac{0}{4} = 0$
Therefore, Jaccard similarity is 0

3. $\mathbf{x} = (0,-1,0,1)$, $\mathbf{y} = (1,0,-1,0)$.

(a) Cosine
Using equation 6
$x.y = 1\times 0+-1\times 0+-1\times 0+1\times 0 = 0$
$||x|| = \sqrt{(0\times 0)+(-1\times -1)+(0\times 0)+(1\times 1)} = \sqrt{2}$
$||y|| = \sqrt{(1\times 1)+(0\times 0)+(-1\times -1)+(0\times 0)} = \sqrt{2}$
Therefore, cosine similarity is 0

(b) Correlation.
Using equation 7,8,9,10
$\overline{x} = 0$
$\overline{y} = 0$
$Cov(x,y) = \frac{(0-0)(1-0)+(-1-0)(0-0)+(0-0)(-1-0)+(1-0)(0-0)}{4} = 0$
$\sigma_x = \sqrt{\frac{(0-0)^2+(-1-0)^2+(0-0)^2+(1-0)^2}{4}} = 0.5$
$\sigma_y = \sqrt{\frac{(1-0)^2+(0-0)^2+(-1-0)^2+(0-0)^2}{4}} = 0.5$
Therefore, correlation is $\frac{0}{0.5\times0.5} = 0$

(c) Euclidean.
Using equation 11
$Euclidean(x,y) = \sqrt{(1-0)^2+(0+1)^2+(-1-0)^2+(0-1)^2} = 2$
Therefore, Euclidean distance is 2

4. $\mathbf{x} = (1,1,0,1,0,1)$, $\mathbf{y} = (1,1,1,0,0,1)$.

(a) Cosine
Using equation 6
$x.y = 1\times1+1\times1+0\times1+1\times0+0\times0+1\times1 = 3$
$||x|| = \sqrt{(1\times1)+(1\times1)+(0\times0)+(1\times1)+(0\times0)+(1\times1)} = 2$
$||y|| = \sqrt{(1\times1)+(1\times1)+(1\times1)+(0\times0)+(0\times0)+(1\times1)} = 2$
Therefore, cosine similarity is $\frac{3}{4}$ i.e 0.75

(b) Correlation.
Using equation 7,8,9,10
$\overline{x} = 0.66$
$\overline{y} = 0.66$
$Cov(x,y) = \frac{(1-0.66)(1-0.66)+(1-0.66)(1-0.66)+(0-0.66)(1-0.66)+(1-0.66)(0-0.66)+(0-0.66)(0-0.66)+(1-0.66)(1-0.66)}{6}$
$Cov(x,y) = \frac{0.33}{6}$
$\sigma_x = \sqrt{\frac{(1-0.66)^2+(1-0.66)^2+(0-0.66)^2+(1-0.66)^2+(0-0.66)^2+(1-0.66)^2}{6}}$
$\sigma_x = \sqrt{\frac{1.3068}{6}}$
$\sigma_y = \sqrt{\frac{(1-0.66)^2+(1-0.66)^2+(1-0.66)^2+(0-0.66)^2+(0-0.66)^2+(1-0.66)^2}{6}}$
$\sigma_y = \sqrt{\frac{1.3068}{6}}$
Therefore, correlation is $\frac{0.33}{1.3068} = 0.25$

(c) Jaccard.
Using equation 12
$Jaccard(x,y) = \sqrt{(1-0)^2+(0+1)^2+(-1-0)^2+(0-1)^2} = 2$
$Jaccard(x,y) = \frac{3}{5} = 0.6$
Therefore, Jaccard similarity is 0.6

5. $\mathbf{x} = (2,-1,0,2,0,-3)$, $\mathbf{y} = (-1,1,-1,0,0,-1)$.

(a) Cosine

Using equation 6

$x.y = 2 \times -1 + -1 \times 1 + 0 \times -1 + 2 \times 0 + 0 \times 0 + (-3 \times -1) = 0$

$||x|| = \sqrt{(2 \times 2) + (-1 \times -1) + (0 \times 0) + (2 \times 2) + (0 \times 0) + (-3 \times -3)} = \sqrt{18}$

$||y|| = \sqrt{(-1 \times -1) + (1 \times 1) + (-1 \times -1) + (0 \times 0) + (0 \times 0) + (-1 \times -1)} = 2$

Therefore, cosine similarity is $\frac{0}{2 \times \sqrt{18}}$ i.e 0

(b) Correlation.

Using equation 7,8,9,10

$\overline{x} = 0$

$\overline{y} = -0.33$

$Cov(x, y) = \frac{(2-0)(-1+0.33)+(-1-0)(1+0.33)+(0-0)(-1+0.33)+(2-0)(0+0.33)+(0-0)(0+0.33)+(-3-0)(-1+0.33)}{6}$

$Cov(x, y) = 0$

$\sigma_x = \sqrt{\frac{(2-0)^2+(-1-0)^2+(0-0)^2+(2-0)^2+(0-0)^2+(-3-0)^2}{6}}$

$\sigma_x = \sqrt{3}$

$\sigma_y = \sqrt{\frac{(-1+0.33)^2+(1+0.33)^2+(-1+0.33)^2+(0+0.33)^2+(0+0.33)^2+(-1+0.33)^2}{6}}$

$\sigma_y = \sqrt{0.55}$

Therefore, correlation is $\frac{0}{\sqrt{3} \times \sqrt{0.55}} = 0$