B565-Data Mining
Homework #1

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

Dr. H. Kurban

Krati Choudhary

January 31, 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.)

(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

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

Solution :

1. Since we are taking absolute value of $x_i - y_i$, it will always be positive which proves the positiveness condition.
2. $|x_i - y_i|$ will always be positive so $x_i - y_i$ and $y_i - x_i$ will give the same value which proves the symmetry condition.
3. $max|x_i - z_i| <= max|x_i - y_i| + max|y_i - z_i|$ which proves the triangle inequality.

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

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

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

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

(c) Suppose $d_0,\ d_1$ are metrics.

i. $d_0 \times d_1$
Solution :
The multiplication of 2 metrics cannot be proven to always be a metric.

ii. $(d_0 + d_1)/(d_0 d_1)$
Solution :
Since $d_0 or d_1$ could be zero which would lead to division by zero condition. Therefore, it cannot be proven to definitely be a metric.

iii. $max\{d_0,\ d_1\}$

iv. Let X be a finite set. Define a distance $d : X \times X \to \mathbb{R}_{\geq 0}$ as $d(x, y) = \frac{||x \cap y||}{||x \cup y|| + 1}$

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.

Problem 2 continued on next page...

R or Python script

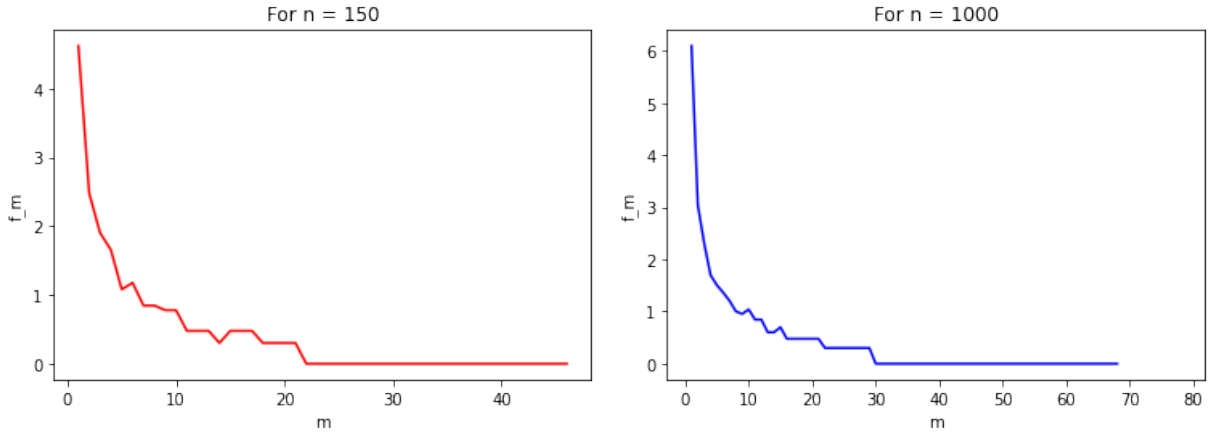
```
# For n = 150
import numpy as np
import matplotlib.pyplot as plt
f_m = []

for m in range(1,101):
    mat = np.random.uniform(0, 1, size=(150,m))
    euc_dist = []
    for j in range(len(mat)-1):
        for k in range(j+1,len(mat)):
            euc_dist.append(np.linalg.norm(mat[j] - mat[k]))
    d_min = min(euc_dist)
    d_max = max(euc_dist)
    d = (d_max-d_min)//d_min
    f_m.append(np.log10(d))

x = np.arange(1, 101)
y = f_m

plt.title("For n = 150")
plt.xlabel("m")
plt.ylabel("f_m")
plt.plot(x, y, color ="red")
```

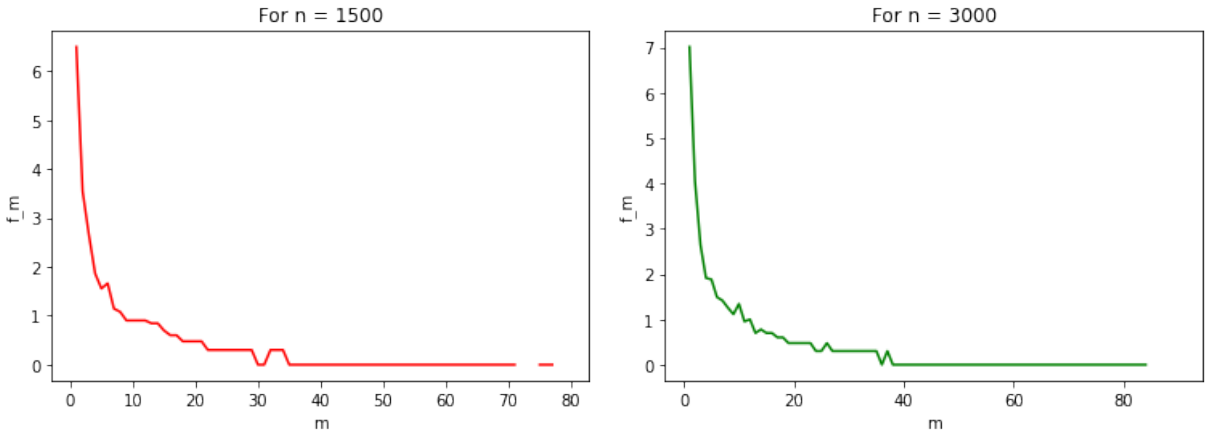
Discussion of Experiments

As we increased the value of n the curve between $f(m)$ and m became smoother with a few outliers.

Plot/s

graphics





Problem 3

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

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

Problem 4

You are data mining 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?
Solution :
To mine this kind of data we can create a record data with attributes like street name, type of areas,

house/apartment number, etc.
Using this data structure we can answer questions such as,

1. How many houses are at a certain street?
2. How many different area types have the same name?
3. Whether an address belongs to an apartment building or a townhouse?

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?
Solution :
The housing data is a tabular collection of data with different attributes describing numerous features of the house like its land size, land contour, year it was built, year remodelled, etc. This data set also has an attributes like the house's sale price, type and condition which can give us an intuition about what the price of a new house that we want to sell would be based on the previous values with corresponding attributes.

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.
   Solution :

   Number of rows is called the number of entries. In the given data set there are 1460 entries.

   ### R/Python script

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

df = pd.read_csv('housing_data.csv')
df.size #df is the dataframe comprising the housing dataset.
```

2. How many unknown or missing data are in the data set? Write the R or Python code in the box below.
   Solution :

   There are a total of 6965 missing values across different attributes that are filled in as NA.

R/Python script

```
# isna().sum() gives the number of missing values in each attribute
df.isna().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.

R/Python script

```
top = df.corr()['SalePrice'].sort_values()
top = top.tail(12)
top = top.iloc[:-1].index[1:]
print(top)

for attr in top:
    plt.scatter(df[attr],df['SalePrice'])
    plt.title(attr)
    plt.xlabel(attr)
    plt.ylabel('Sale Price')
    plt.show()
```

Discussion of Findings

The top 10 attributes that affect the value of Sale Price are:
YearRemodAdd, YearBuilt, TotRmsAbvGrd, FullBath, 1stFlrSF, TotalBsmtSF, GarageArea, GarageCars, GrLivArea, OverallQual
All these attribute affect the Sale Price in different ways and were found using correlation between these attributes and Sale Price. A scatter plot is a good way to demonstrate these trends as we have 1460 different values of SalePrice that range between 0-700000 with a few outliers.

(a) **1stFlrSF :** Most of the values of this attribute are concentrated in the range 0 to around 2000 and the corresponding SalePrice is in the range 0 to 500000. This shows there isn't much variance in the values.

(b) **FullBath :** SalePrice value increase as the number of full bathrooms increase but not that much.

(c) **GarageArea :** The houses with no Garage area have lowest SalePrice values. On the other hand, ones that do have a well distributed SalePrice value so this attribute does not give us much information.

(d) **GarageCars :** The visualization shows that most people have 3 cars so the houses with car space for 3 cars is higher than others.

(e) **GrLivArea :** The data points are pretty dense and and the SalePrice does not increase over 500000 for most houses.

(f) **OverallQual :** The overall quality of the house highly affects the SalePrice. As the index increases, the house price also increases significantly.

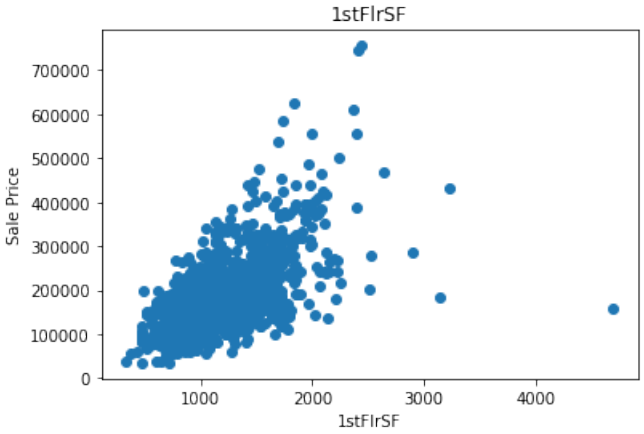
(g) **TotalBsmtSF :** Total square feet of basement area significantly increases the house SalePrice.

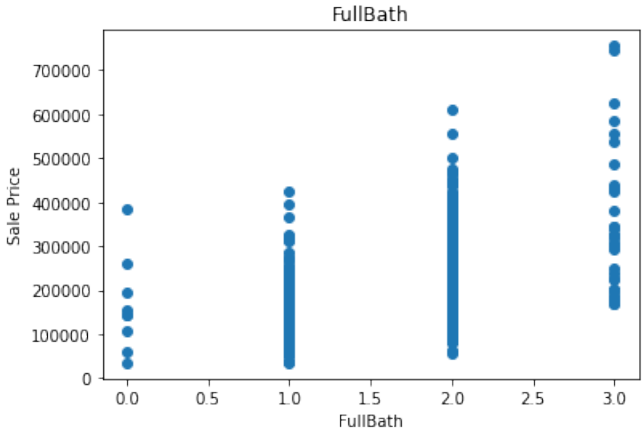
(h) **TotRmsAbvGrd :** As the number of rooms above ground in the house increase, the cost of house also increases.

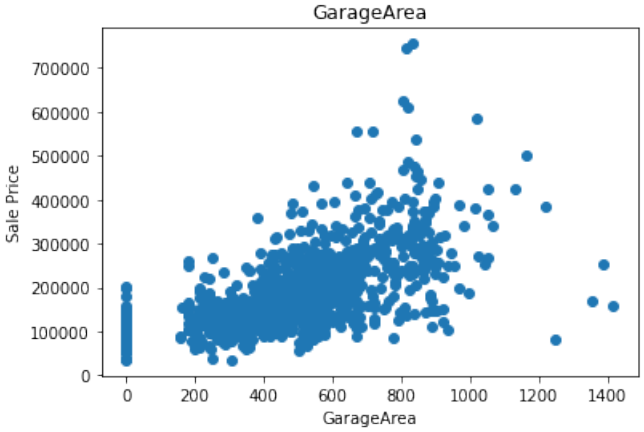
(i) **YearBuilt :** Newer houses cost more than the ones that are older.

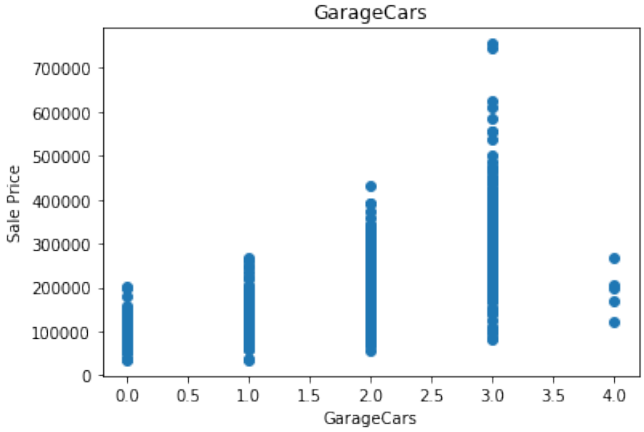
(j) **YearRemodAdd :** The year a house was remodeled does not significantly increase the cost of the house.

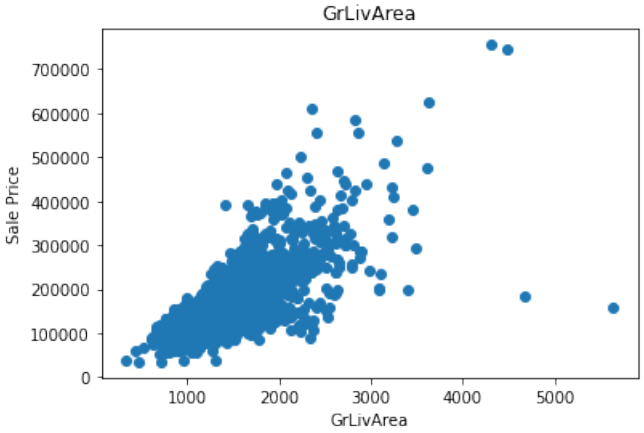
Plot/s

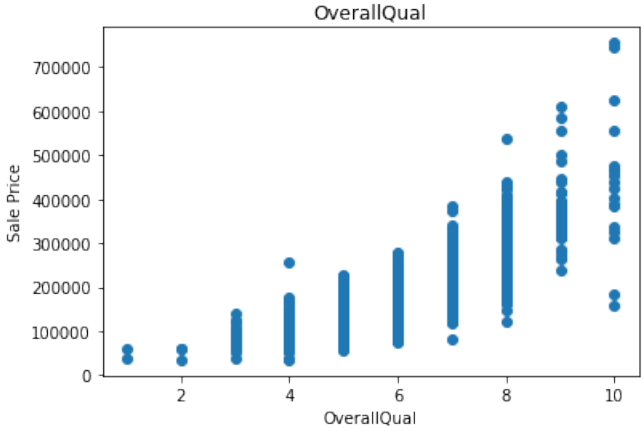


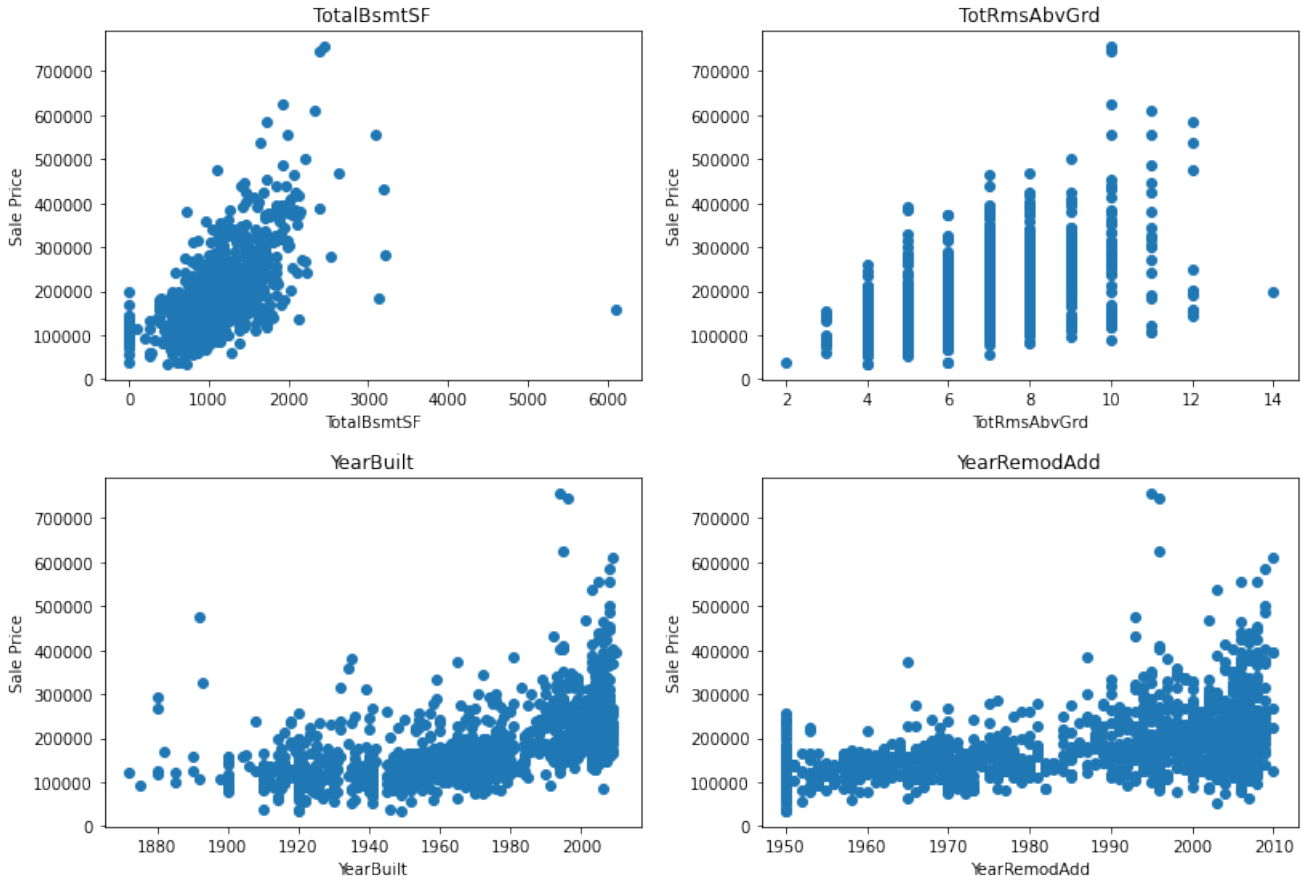












4. Make a histogram/bar plot for each of those 10 attributes influencing `SalePrice` and discuss the distribution of values, *e.g.*, are uniform, skewed, normal of those attributes. Place images of these histograms into the document. Show the R/Python code that you used below and discussion below that.

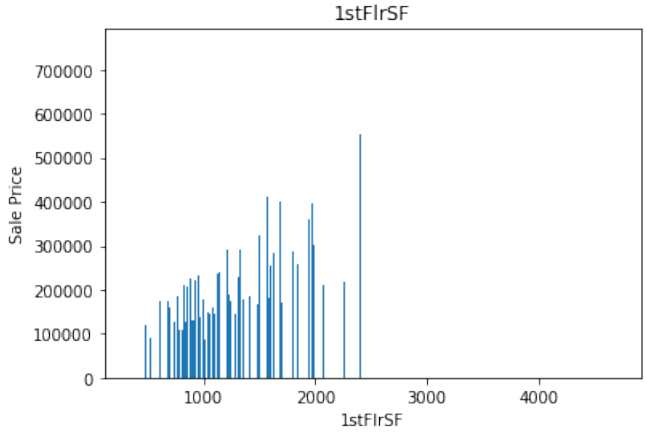
Python script

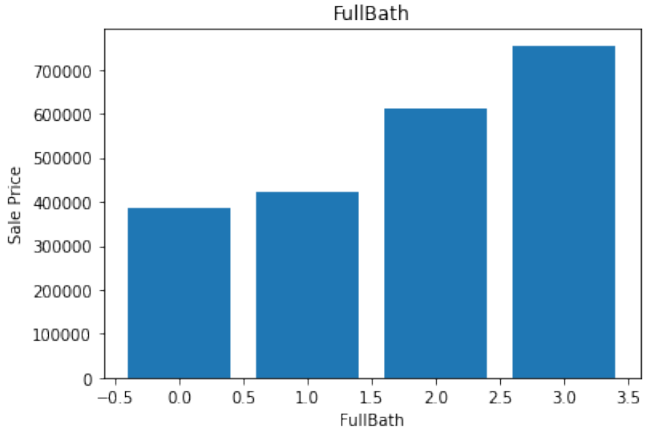
```
#Bar plot
for attr in top:
    plt.bar(df[attr],df['SalePrice'])
    plt.title(attr)
    plt.xlabel(attr)
    plt.ylabel('Sale Price')
    plt.show()
```

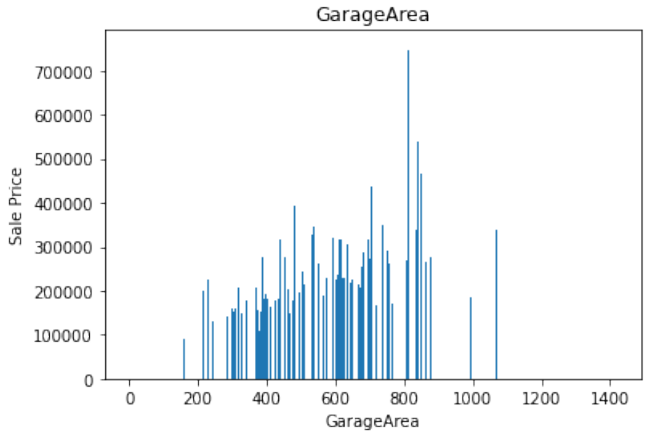
Discussion of Attributes

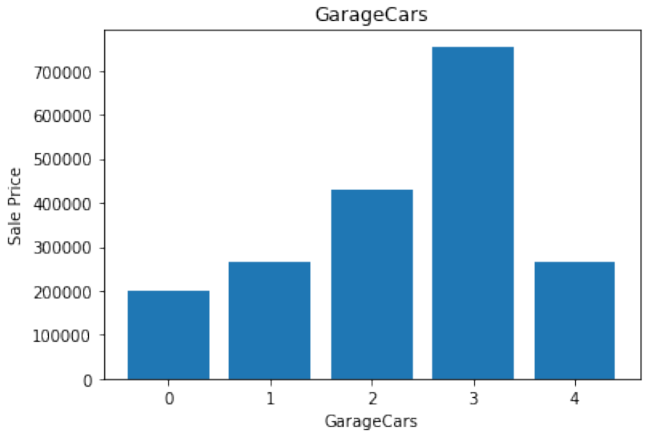
Almost all attributes that affect the SalePrice are pretty right skewed which implies that as the value of these attributes increases the SalePrice value also increases.

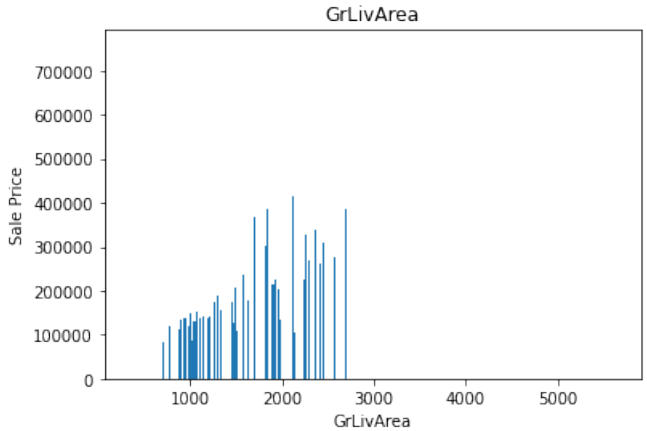
Histograms/Bar Plots

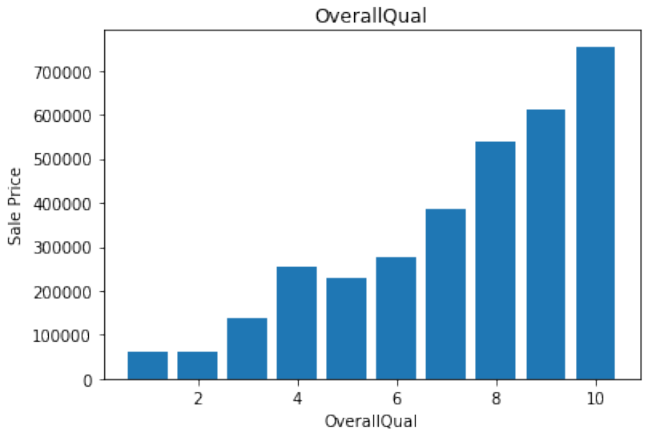


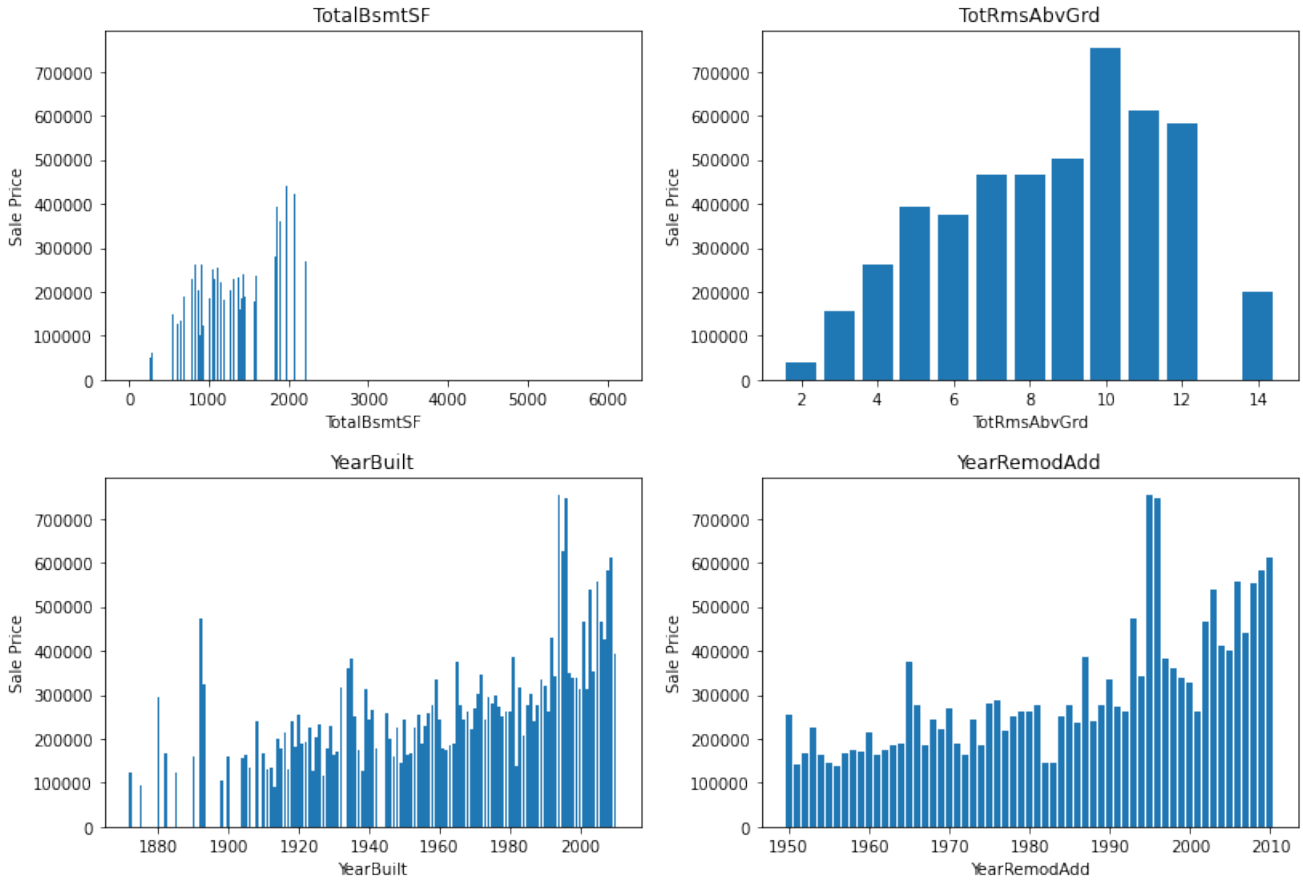












Discussion of simply removing tuples

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

Solution :

If we simply remove the tuples containing missing or unknown values, we might loose some important data and create bias in the model, which in turn would affect our end decision.

Problem 6

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

1. Is noise ever interesting or desirable? Outliers?
2. Can noise objects be outliers?
3. Are noise objects always outliers?
4. Are outliers always noise objects?
5. Can noise make a typical value into an unusual one, or vice versa?

Solution :

1. Noise is never interesting or needed as it distorts the original data. On the other hand, outliers might contain some interesting trend data which could help identify abnormalities in some problems.

2. Noise objects can be outliers as noise can make data abnormal which might get detected as an outlier.

3. No, noise data is not always outlier as some noise might be detected in the range of the normal data.

4. No, outliers are not always noise objects as outliers can just be a genuine unique case that occurred while collecting data.

5. Noise can make some normal values to appear as outliers and at the same time might make some outliers getting classified as normal values.

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.
   Solution :
   Selecting n elements this way would help us get few elements from each of the groups which would be important when each group is divided based on some category like genders.

2. We randomly select n elements from the data set, without regard for the group to which an object belongs.
   Solution:
   Selecting n elements like this might cause some important data getting missed which would cause bias in further processing and while training a model on that data. The biggest example of such a blunder was Google Photos labelling people of certain race and skin color as gorillas.

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?
   Solution :
   If a term occurs in just one document, this transformation will determine how important that word is for that document and help identify the document from a group based on its relevance. But if the term occurs in every document then this transformation will determine that this word is not important in the document distinguishing query.

2. What might be the purpose of this transformation?
   Solution :
   This transformation helps reduce the weight of common words so that we can focus on the more important terms which would be more helpful in finding the relevant document.

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.

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

   Solution :
   Hamming Distance = 3
   Jaccard Similarity = $2/(2+1+2) = 2/5$

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.)

   Solution :
   Hamming distance is more similar to SMC as both of them consider all bits in both the binary vectors during calculation. On the other hand, Jaccard is more similar to cosine measure as both of them can only be of the range 0 to 1.

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.)

   Solution :
   Jaccard similarity would be a better measure to calculate how similar two organisms of different species are, as it counts the number of bits that have same values at the same place in different vectors by ignoring the bits that are 0 in both since if a certain gene is not present in both it is not relevant for their comparison.

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.)

   Solution :
   Hamming distance can be used to compare the genetic makeup of two organisms of same species since they share the same genes for the most part. If we try to compare similarity measure computed by Jaccard or any other measure of similarity we won't get very useful information.

Problem 10

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

Problem 10 continued on next page...

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.

Solution :

1.
   - **Cosine similarity** $= \frac{\mathbf{x}\cdot\mathbf{y}}{\sqrt{\mathbf{x}\cdot\mathbf{x}}\sqrt{\mathbf{y}\cdot\mathbf{y}}} = \frac{1*2+1*2+1*2+1*2}{\sqrt{1*1+1*1+1*1+1*1}*\sqrt{2*2+2*2+2*2+2*2}} = \frac{8}{2*4} = 1$
   - **Correlation** $= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$

     Since, $\bar{x} = 1$ and $\bar{y} = 2$ therefore, $(x_i - \bar{x}) = 0$ and $(y_i - \bar{y}) = 0$
     So, correlation $= 0$
   - **Euclidean** $= d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} = \sqrt{(1-2)^2 + (1-2)^2 + (1-2)^2 + (1-2)^2}$

     $= \sqrt{(-1)^2 + (-1)^2 + (-1)^2 + (-1)^2} = \sqrt{1+1+1+1} = \sqrt{4} = 2$
2.
   - **Cosine similarity** $= \frac{\mathbf{x}\cdot\mathbf{y}}{\sqrt{\mathbf{x}\cdot\mathbf{x}}\sqrt{\mathbf{y}\cdot\mathbf{y}}} = \frac{0*1+1*0+0*1+1*0}{\sqrt{0*0+1*1+0*0+1*1}*\sqrt{1*1+0*0+1*1+0*0}} = \frac{0}{\sqrt{2}*\sqrt{2}} = 0$
   - **Correlation** $= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$

     $\bar{x} = 1/2$ and $\bar{y} = 1/2$

     $= \frac{(0-1/2)(1-1/2)+(1-1/2)(0-1/2)+(0-1/2)(1-1/2)+(1-1/2)(0-1/2)}{\sqrt{(0-1/2)^2+(1-1/2)^2+(0-1/2)^2+(1-1/2)^2}*\sqrt{(1-1/2)^2+(0-1/2)^2+(1-1/2)^2+(0-1/2)^2}}$

     $= \frac{-1/4-1/4-1/4-1/4}{\sqrt{1/4+1/4+1/4+1/4}*\sqrt{1/4+1/4+1/4+1/4}}$

     $= \frac{-1}{1} = -1$
   - **Euclidean** $= d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} = \sqrt{(0-1)^2 + (1-0)^2 + (0-1)^2 + (1-0)^2}$

     $= \sqrt{(-1)^2 + (1)^2 + (-1)^2 + (1)^2} = \sqrt{1+1+1+1} = \sqrt{4} = 2$
   - **Jaccard** $= \frac{M_{11}}{M_{10}+M_{01}+M_{11}}$

     $M_{11} = 0, M_{10} = 2, M_{01} = 2$

     $J(x, y) = \frac{0}{2+2+0} = 0$
3.
   - **Cosine similarity** $= \frac{\mathbf{x}\cdot\mathbf{y}}{\sqrt{\mathbf{x}\cdot\mathbf{x}}\sqrt{\mathbf{y}\cdot\mathbf{y}}}$
     $Since, \mathbf{x}\cdot\mathbf{y} = 0*1 + (-1)*0 + 0*(-1) + 1*0 = 0$
     $cosine = 0$
   - **Correlation** $= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$

     Since, $\bar{x} = 0$ and $\bar{y} = 0$ therefore, $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = 0$
     So, correlation $= 0$
   - **Euclidean** $= d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} = \sqrt{(0-1)^2 + (-1-0)^2 + (0-(-1))^2 + (1-0)^2}$

     $= \sqrt{(-1)^2 + (-1)^2 + (1)^2 + (1)^2} = \sqrt{1+1+1+1} = \sqrt{4} = 2$

4. • **Cosine similarity** $= \frac{\mathbf{x}\cdot\mathbf{y}}{\sqrt{\mathbf{x}\cdot\mathbf{x}}\sqrt{\mathbf{y}\cdot\mathbf{y}}} = \frac{1*1+1*1+0*1+1*0+0*0+1*1}{\sqrt{1*1+1*1+0*0+1*1+0*0+1*1}*\sqrt{1*1+1*1+1*1+0*0+0*0+1*1}} = \frac{1+1+0+0+0+1}{\sqrt{4}*\sqrt{4}}$
$= 3/4$

• **Correlation** $= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$

$\bar{x} = 2/3$ and $\bar{y} = 2/3$

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

$= \frac{1/9+1/9+(-2/9)+(-2/9)+4/9+1/9}{\sqrt{1/9+1/9+4/9+1/9+4/9+1/9}*\sqrt{1/9+1/9+1/9+4/9+4/9+1/9}}$

$= \frac{1/3}{\sqrt{4/3}*\sqrt{(4/3)}} = \frac{1/3}{4/3} = 1/4$

• **Jaccard** $= \frac{M_{11}}{M_{10}+M_{01}+M_{11}}$

$M_{11} = 3, M_{10} = 1, M_{01} = 1$

$J(x,y) = \frac{3}{1+1+3} = 3/5$

5. • **Cosine similarity** $= \frac{\mathbf{x}\cdot\mathbf{y}}{\sqrt{\mathbf{x}\cdot\mathbf{x}}\sqrt{\mathbf{y}\cdot\mathbf{y}}}$

$= \frac{2*(-1)+(-1)*1+0*(-1)+2*0+0*0+(-3)*(-1)}{\sqrt{2*2+(-1)*(-1)+0*0+2*2+0*0+(-3)*(-3)}*\sqrt{(-1)*(-1)+1*1+(-1)*(-1)+0*0+0*0+(-1)(-1)}}$

$= \frac{-2-1+0+0+0+3}{\sqrt{4+1+0+4+0+9}*\sqrt{1+1+1+0+0+1}} = \frac{0}{3*\sqrt{2}*2} = 0$

• **Correlation** $= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$

$\bar{x} = 0$ and $\bar{y} = -1/3$

$= \frac{2*(-2)/3+(-1)*4/3+0+2*1/3+0+(-3)*(-2)/3}{\sqrt{4+1+0+4+0+9}*\sqrt{4/9+16/9+4/9+1/9+1/9+4/9}}$

$= \frac{-4/3-4/3+2/3+6/3}{\sqrt{18}*\sqrt{(30/9)}} = \frac{0}{2\sqrt{15}} = 0$