DS - 2

Q.1. Why do we need to preprocess the data?

→ 1. Data Preprocessing is the labor-intensive phase that covers all aspects of preparing the final data set, which shall be used for subsequent phases, from the initial, raw, dirty data.

2. The data may contain -
- fields that are redundant
- missing values
- outliers
- data in a form not suitable for the data mining models.
- values not consistent with policy or common sense.

Hence, it is necessary to preprocess the data.

3. Depending on data set, data preprocessing alone can account for 10-60% of all time & effort for the entire data science process.

4. Data preprocessing involves -

i] Data preparation -
- Evaluate the quality of data.
- Clean the raw data.
- Deal with missing data.
- Perform transformations on certain variables

a] Data understanding
- Exploratory data analysis (EDA).

Q.2. Describe the possible negative effects of proceeding directly to mine data that has not been preprocessed?

→ 1. fields are redundant
2. missing values
3. outliers
4. data in a form not suitable for data mining models.
5. values not consistent with policy or common sense.

Q.3. What are four ways to handle missing data in dataset? Of the four methods for handling missing data, which method is preferred?

→

Page No.
Date

Q.4. What is an outlier? Why do we need to treat outliers carefully?

→ - Outliers are extreme values that go against the trend of remaining data.

- An outlier is an observation point that is distant from other observations.

- In data mining, outlier detection is the identification of rare items, events or observations which raise suspicions by differing significantly from the majority of data.

- Identifying outliers is imp because they may represent errors in data entry.

- Statistical methods are sensitive to presence of outliers and may deliver unreliable results.

- Method for identifying outliers for numeric variables is to examine a histogram of the variable.

Q.5. Explain graphical methods for identifying errors.

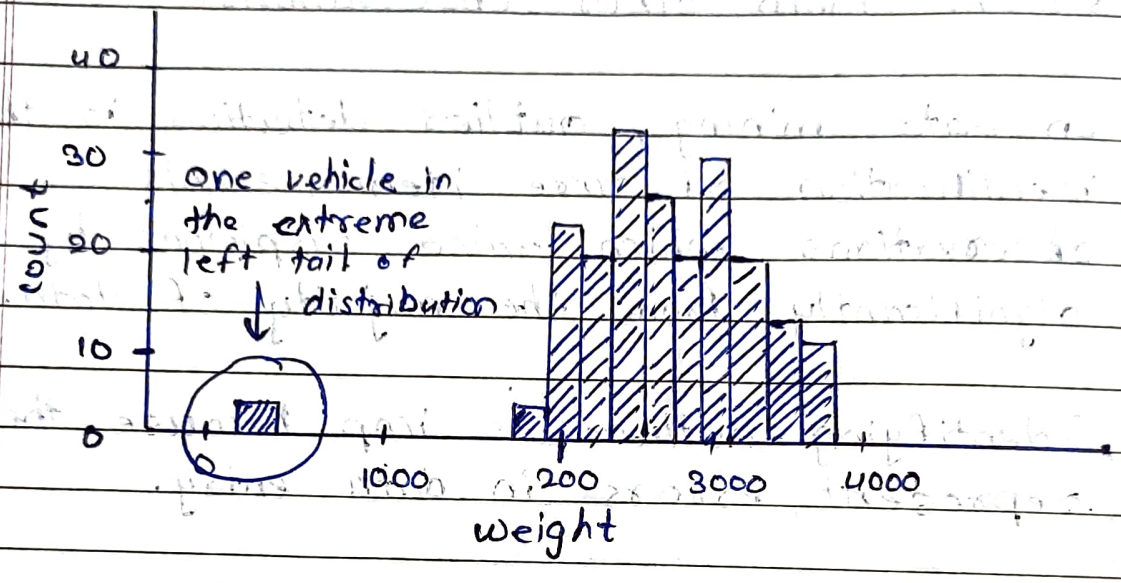
→ 1) histogram -

Method for identifying outliers for numeric variables is to examine a histogram of the variable.

In histogram, existence of outliers can be detected by isolated bars.

Histograms are generally used in univariate settings where we graph the data distribution of a single variable & identify outliers that falls ~~out~~ outside of data distribution.

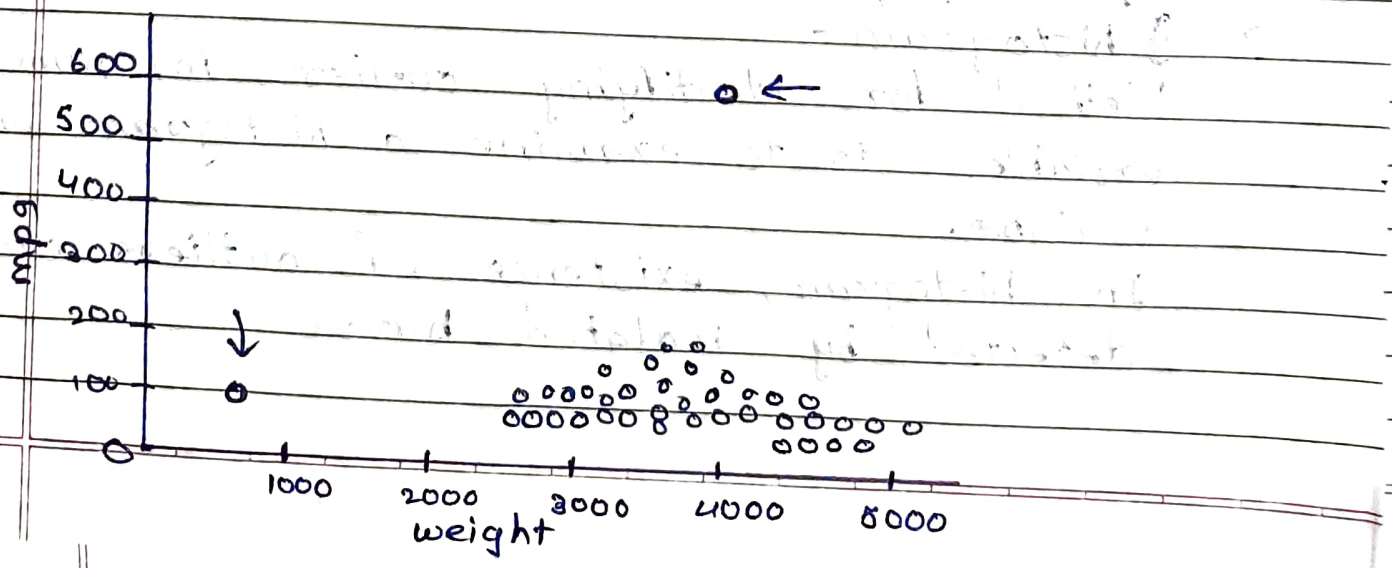
Eg -



2] ~~Sa~~ ~~Sc~~ Scatterplots -

Scatterplots can be explicitly detect when a dataset or ~~particul~~ particular feature contains outliers.

Sometimes two-dimensional scatter plots can help to reveal outliers in more than one variable.



Q.6. Explain measures of center & spread.

→ - Measures of center are a special case of measures of location.

- Commonly used numerical measures of center are mean, median & mode.

- Measures of location are percentiles & quantiles.

- Mean of a variable is simply the avg of the valid values taken by the variable.

- For variables that are not extremely skewed, the mean is usually not too far from the variable center.

- For extremely skewed datasets, the mean becomes less representative of the variable center.

- Mean is sensitive to presence of outliers.

- Analysts sometimes prefer median as ~~the~~ alternative measures of center.

- Median is resistant to presence of outliers.

- Other analysts may prefer to use mode.

- Measures of center do not always concur as to where the center of data set lies.

- Measures of location are not sufficient to summarize a variable effectively.

- Two variables may have the very same values for the mean, median & mode & yet have different natures.

Ex -

portfolio A	portfolio B
1	7
11	8
11	11
11	11
16	13

The mean P/E ratio is 10, median is 11 & the mode is 11 for each portfolio.
These measures of center do not provide us with a complete picture.
Portfolio A's P/E ratios are more spread out than those of portfolio B. So the measures of variability for portfolio A should be larger than thos of B.

- Typical measures of ~~sensitivity~~ variability include range [maximum - minimum], the deviation [SD], mean absolute deviation & the interquartile range [IQR].

- $$SD = \sqrt{\frac{\Sigma (x - \bar{x})^2}{n-1}}$$

SD is sensitive to presence of ~~outl~~ outliers.

Mean absolute deviation,

$$MAD = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

Q.7. Explain why data analysts need to normalize their numeric variables.

→ - Variables tend to have ranges that vary greatly from each other.

- Ex - ① players' batting averages will range from zero to less than 0.400.
② number of home runs hit in season will range from 0 to 70.

- Such differences in ranges will lead to a tendency for the variable with greater range to have undue influence on the results.

- Hence data miners/ data scientists should normalize their numeric variables.

- Neural networks benefit from normalization.

- Benefit to algorithms that make use of distance measures, such as the k-nearest neighbours algorithm.

Q. 8. Explain min-max normalization, z-score standardization & decimal scaling data transformation techniques.

1] Min-Max normalization.

Let X refer to our original field value, and x^* refer to normalized field value.

Min-Max normalization works by seeing how much greater the field value is than the minimum value min(x) and scaling this difference by range.

$$x^*_{mm} = \frac{X - min(x)}{range(x)} = \frac{X - min(x)}{max(x) - min(x)}$$

Ex -

mean	3005.490
min	1613
max	4997
range	3384
SD	852.646

For vehicle weighing 1613 pounds, the min-max normalization is:-

$$x^*_{mm} = \frac{1613 - 1613}{3384} = 0.$$

2] Z-score standardization

Z-score standardization is very widespread in world of statistical analysis.

It works by taking difference bet^n the field value & field mean value & scaling this difference by SD.

$$z\text{-score} = \frac{X - mean(x)}{SD}$$

Page No.
Date

Ex - mean 3005.490
min 1613
max 4997
range 3384
SD 852.646.

For the avg weight vehicle, z-score is.

$$z\text{-score} = \frac{x - \text{mean}(x)}{SD} = \frac{3005.490 - 3005.490}{852.646}$$

$$= 0.$$

3) Decimal scaling.

Decimal scaling ensures that every normalized value lies bet^n -1 & 1.

$$x^* \text{decimal} = \frac{x}{10^d}$$

where d represents the number of digits in the data value with the largest absolute value.

Ex - The decimal scaling for the min & max weight are

$$x^* \text{decimal} = \frac{1613}{10^4} = 0.1613$$

$$x^* \text{decimal} = \frac{4997}{10^4} = 0.4997.$$

Page No.
Date

Q.9. For the stock price data given below, find min-max normalization stock price for all the stock prices.

10	7	20	12	75	15	9	18	4	12	8	14

→

The given dataset is,
10, 7, 20, 12, 75, 15, 9, 18, 4, 12, 8, 14.

min-max normalization $= x^*_{mm} = \frac{x - \min(x)}{\text{range}(x)}$

here, $\min(x) = 4$

$\text{range}(x) = \max(x) - \min(x) = 75 - 4 = 71.$

$\therefore x^*_{mm}(10) = \frac{10-4}{71} = 0.0845$

$x^*_{mm}(7) = \frac{7-4}{71} = 0.0422$

$x^*_{mm}(20) = \frac{20-4}{71} = 0.2253$

$x^*_{mm}(12) = \frac{12-4}{71} = 0.1126$

$x^*_{mm}(75) = \frac{75-4}{71} = 1$

$x^*_{mm}(15) = \frac{15-4}{71} = 0.1549$

$x^*_{mm}(9) = \frac{9-4}{71} = 0.0704$

$x^*_{mm}(18) = \frac{18-4}{71} = 0.1971$

$x^*_{mm}(4) = \frac{4-4}{71} = 0$

$x^*mm(12) = \frac{12-4}{71} = 0.1126$

$x^*mm(8) = \frac{8-4}{71} = 0.0563$

$x^*mm(14) = \frac{14-4}{71} = 0.1408.$

Q.10. For the stock price data given below, find z-score standardization stock price for all the stock prices.

10 7 20 12 75 15 9 18 4 12 8 84

→

The given dataset is
10, 7, 20, 12, 75, 15, 9, 18, 4, 12, 8, 84.

$$z\text{-score} = \frac{x - mean(x)}{SD(x)}$$

here, $mean(x) = \frac{10+7+20+12+75+15+9+18+4+12+8+84}{12}$

$mean(x) = 16.16$

$$SD(x) = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$$

$$= \sqrt{\frac{(10-16.16)^2 + (7-16.16)^2 + (20-16.16)^2 + (12-16.16)^2 + (75-16.16)^2 + (15-16.16)^2 + (9-16.16)^2 + (18-16.16)^2 + (4-16.16)^2 + (12-16.16)^2 + (8-16.16)^2 + (4-16.16)^2}{12-1}}$$

$$= \sqrt{\frac{4051.71}{11}}$$

SD(x). = 19.19

$$\therefore \text{Z-score}(10) = \frac{10 - 16.16}{19.19}$$

$$= -0.32$$

Z-score (7) = -0.48
Z-score (20) = 0.20
Z-score (12) = -0.22
Z-score (75) = 3.07
Z-score (15) = -0.06
Z-score (9) = -0.37
Z-score (18) = 0.10
Z-score (4) = -0.63
Z-score (12) = -0.22
Z-score (8) = -0.43
Z-score (4) = -0.63

Q. 11. For the stock price data given below, find the decimal scaling stock price for all the stock prices.
10, 7, 20, 12, 75, 15, 9, 18, 4, 12, 8, 14.

→

The given dataset is:
10, 7, 20, 12, 75, 15, 9, 18, 4, 12, 8, 14.

$$X^*_{decimal} = \frac{X}{10^d}$$

Page No.
Date

x^* decimal (10) $= \frac{10}{10^2} = 0.1$

x^* decimal (7) $= \frac{7}{10^1} = 0.7$

x^* decimal (20) $= \frac{20}{10^2} = 0.20$

x^* decimal (12) $= \frac{12}{10^2} = 0.12$

x^* decimal (75) $= \frac{75}{10^2} = 0.75$

x^* decimal (15) $= \frac{15}{10^2} = 0.15$

x^* decimal (9) $= \frac{9}{10^1} = 0.9$

x^* decimal (18) $= \frac{18}{10^2} = 0.18$

x^* decimal (4) $= \frac{4}{10^1} = 0.4$

x^* decimal (12) $= \frac{12}{10^2} = 0.12$

x^* decimal (8) $= \frac{8}{10^1} = 0.8$

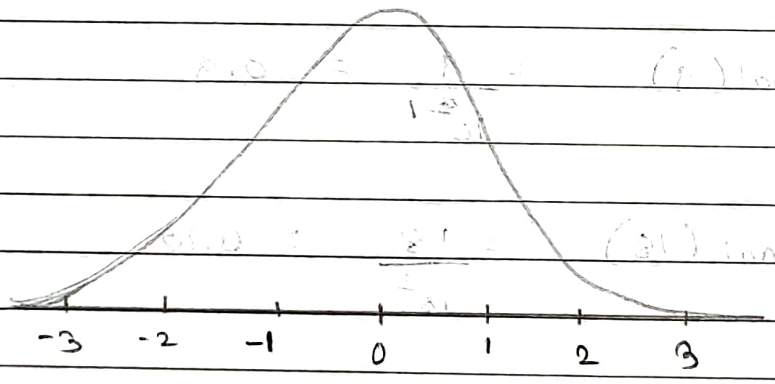
x^* decimal $(14) = \frac{14}{10^2} = 0.14$

Q.12. Explain transformations that can be applied to achieve normalization in the data? Why normalized dataset is preferred?

→ - The normal distribution is a continuous probability distribution commonly known as bell curve, which is symmetric.

- A distribution is symmetric if it looks the same to the left and right of center point.

- Normal distribution that has mean $\mu = 0$ & SD $\sigma = 1$ known as the standard normal distribution Z.

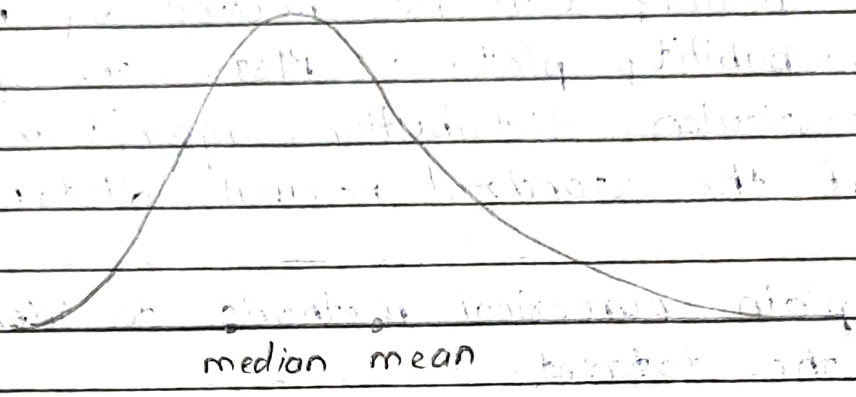


standard normal distribution.

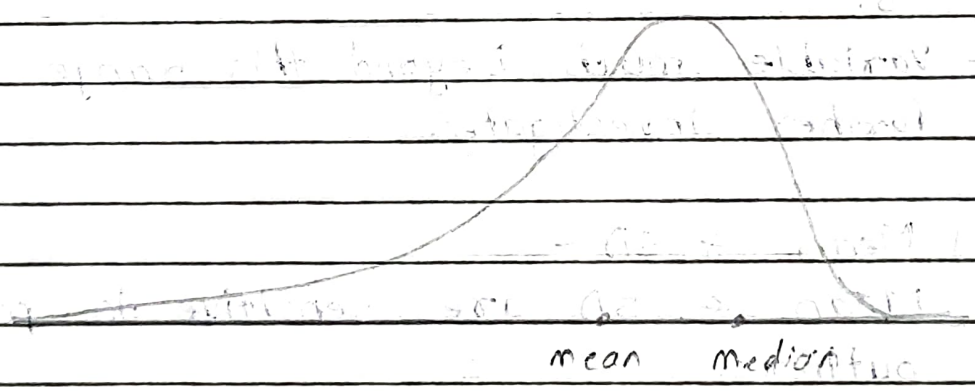
- Z-standardized data will have mean = 0 & SD = 1 but the distribution may still be skewed.

- skewness = $\frac{3(\text{mean} - \text{median})}{\text{standard deviation}}$

- Right skewed data has positive skewness. Mean is greater than median.



- Left skewed data has negative skewness. Mean is smaller than median.



- For perfectly symmetric data, the mean, median, mode are all equal & so the skewness equals zero.

- Z-score standardization has no effect on skewness.

Page No.
Date

12. Common transformations are -
- natural log transformation
- square root transformation
- inverse square root transformation.

→ - Normality can be checked by normal probability plot. → Plots the quantiles of particular distribution against the quantiles of the standard normal distribution.

Q.13. Explain numerical methods to identify outliers in the dataset.

→ 1] z-score
- The z-score method states that a data value is an outlier if it has a z-score that is either less than -3 or greater than 3.
- Variable much beyond this range may be further investigated.

2] Mean & SD -
- Mean & SD are sensitive to presence of outliers.
- They are part of formula ~~of~~ for z-score

3] Interquartile Range (IQR) -

- IQR is robust statistical method for outlier detection, which are less sensitive to the presence of outliers.
- Quartiles of data set divide data set in four parts each containing 25% of data.

First quartile (Q_1) - 25^{th} percentile
second quartile (Q_2) - 50^{th} percentile, (median)
third quartile (Q_3) - 75^{th} percentile

- IQR is calculated as $IQR = Q_3 - Q_1$.

- A data value is an outlier if
  - it is located 1.5(IQR) or more below Q_1
  - it is located 1.5(IQR) or more above Q_3.

- Ex - $Q_1 = 70$ $Q_3 = 80$
  $\therefore$ $IQR = 80 - 70 = 10$
  A test score would be robustly identified as outlier if -
  - it is lower than
    $Q_1 - 1.5(IQR) = 70 - 1.5(10) = 55$.
  - it is higher than
    $Q_3 + 1.5(IQR) = 80 + 1.5(10) = 95$.

Q.16. What is flag variable? What is its use?

→ - A flag variable which is also called as dummy variable / indicator variable is a a categorical variable taking only two values 0 & 1.

- Some analytical methods, such as regression require predictors to be numeric.

- Need to recode the categorical variable into one or more flag variables.

- Eg - The categorical predictor gender, taking values for female & male could be recorded into flag variable gender_flag as follows:-
if gender = female then gender_flag = 0;
if gender = male then gender_flag = 1;

- When a categorical predictor takes $k \geq 3$ possible values, then define $k-1$ dummy variables and use the unassigned category as the reference category.

- Ex - region has four possible categories [~~north~~ north, south, east, west], then the analyst could define following $k-1 = 3$ flag variables:

north_flag : If region = north then north_flag = 1;
otherwise north_flag = 0;
east_flag : If region = east then east_flag = 1;
otherwise east_flag = 0;

south_flag: If region = south then south_flag = 1;
otherwise south_flag = 0;

Flag variable for west is not needed, as region = west is already uniquely identified by zero values for each of the three existing flag variables.

Q.17. Explain techniques for binning numerical variables.

→ 1] Equal width binning -
It is not recommended, as the width of categories can be greatly affected by the presence of outliers.
Ex - $X = \{1, 1, 1, 1, 1, 2, 2, 11, 11, 12, 12, 44\}$ & $k = 3$.

low: $0 \leq x \leq 15$ - which contains all the data values except one
medium: $15 \leq x \leq 30$, - which contains no data values at all
high: $30 \leq x < 45$, which contains a single outlier.

2] Equal frequency binning -
$X = \{1, 1, 1, 1, 1, 2, 2, 11, 11, 12, 12, 44\}$.
we have $n = 12$, $k = 3$ & $n/k = 4$.
low: contains first four data values, all $x = 1$.
medium: contains next four data values $\{1, 2, 2, 11\}$.
high: contains last four data values $\{11, 12, 12, 44\}$.

3] Binning by clustering.

4] Binning based on predictive value.

Q.20. How to remove duplicate records from dataset?

→ - Records may have been inadvertently copied, thus creating duplicate records.

- Duplicate records lead to an overweighting of data values.

- Only one set of them should be retained.
- For ex, if the ID field is duplicated, then definitely remove the duplicate records.

- Data analysts should apply common sense.

- Removing duplicate records is not particularly difficult.

- Most statistical packages & database systems have built-in commands that group records together.

- In database language SQL, this command is called GroupBy.

Page No.
Date

Q.18. Explain why we might not want to remove a variable that had 90% or more missing values.

→ - There is a common practice to remove variables for which 90% or more of the values are missing.

- But ~~it~~ there may be pattern in the missingness and therefore useful information.

- Challenge to any strategy for imputation of missing data.

- If 10% of data is representative, then choose to imputation of missing 90%.

- Imputation is based on regression & or decision tree methods.

- Hence avoid removing variables having missing values.

Q.19. Explain why we might not want to remove a variable just because it is highly correlated with another variable.

→ - There is a common practice to remove variables that are strongly correlated.

- Inclusion of correlated variables may
  - at best double-count a particular aspect of the analysis

- at worst lead to instability of model results.

- Removing one of the variables might discard imp ~~if~~ info.

- Hence avoid removing one of the correlated variables.

- ~~&~~ Principal components analysis may be applied, where the common variability in correlated predictors may be translated into a set of uncorrelated principal components.

Q.15. For the stock price data given below, identify all possible stock prices that would be outliers using a) z-score method
b) IQR method.

10, 7, 20, 12, 75, 15, 9, 18, 4, 12, 8, 14.

→ a] z-score method -

z-score method states that a data value is outlier if it has a z-score that is either less than -3 or greater than 3.

[calculating z-score as in Q.10].

Hence, 75 is outlier as the z-score for it is 3.07.