**Homework 2 - Kanika Yadav CWID- 20009527 Date – 09/22/2022**

Answer questions 3, 6, 7, 8, 9, 10, 12, 13, 14, and 15 in the Exercises section from Chapter 2 of the Larose textbook.

In your submission, you must:

* Clearly answer each question in a Word document and submit the Word document to the Homework 2 assignment.
* Write your answers completely and elaborate on your ideas as much as possible.

--------------------------------------------------------------------------------------------------------------------------

Q.3 Explain why zip codes should be considered text variables rather than numeric.

**Ans.**

- Zip codes differ from country to country  
- if we stick them to be numeric then we might face issues with alphanumeric codes  
- Format won’t be consistent  
- Data won’t be suitable to feed into the data mining model to process  
- important to have consistency and reusability of values   
- To save further data cleaning and data transformation steps that would cost more it is wiser plan to keep the Zip code as Text Variables than numeric.

Q.6 . True or false: All things being equal, more information is almost always better.

**Ans.** True.

We can get more insights from underlying relationship of given area of discourse from the data.

Q.7 Explain why it is not recommended, as a strategy for dealing with missing data, to simply omit the records or fields with missing values from the analysis.

**Ans.**

Not recommended to simply omit the records or fields with missing values from the analysis –   
- the pattern of missing values may be systematic   
- simply deleting the records with missing values would lead to biased subset of the data  
- Omitting the records just because it has missing values might omit other important information related in all the other fields  
- Almost 80% of the records would have at least one missing value, and it is recommended to not to lose on important information from various other fields related to dame data row.

Q.8. Which of the four methods for handling missing data would tend to lead to an under estimate of the spread (e.g., standard deviation) of the variable? What are some benefits to this method?   
**Ans.**

Out of the four methods for handling missing data, replacing data with Mean value of the column will lead to an underestimate of the spread of the variable.   
Some benefits of this replacing missing values with mean –

* Mean can be generated by a software
* Hence a software can take care of missing data issues to fill the holes in the data
* Instead of using constants replacing the missing data with mean can be called as gamble and its benefits must be weighed against possible invalidity of the results.

Q.9. What are some of the benefits and drawbacks for the method for handling missing data that chooses values at random from the variable distribution?

**Ans.**Benefits of method for handling missing data that choose values at random from variable distribution–  
- measures of centre and spread will remain to the original  
- there will not much deviation from the expected results  
  
Drawbacks  
- no guarantee that resulting record make sense   
- resulting record might not exist in real  
  
Hence we need data imputation methods.

Q.10. Of the four methods for handling missing data, which method is preferred?   
**Ans.** Out of the four methods of handling missing data the method of replacing missing values with imputed values that take of knowledge about the missing value. It helps to be more realistic and most likely missing value considering all the other attributes for a particular record.

This process can be formulated by the tools like multiple regression or classification and regression trees.

Q.12. Make up a data set, consisting of the heights and weights of six children, in which one of the children is an outlier with respect to one of the variables, but not the other. Then alter this data set so that the child is an outlier with respect to both variables.   
**Ans.**

**Table 1.** Data for child – 3 is an outlier with respect to one of the variables

|  |  |  |
| --- | --- | --- |
| **Child** | **Height(cms)** | **Weight(lbs)** |
| 1 | 160 | 110 |
| 2 | 158 | 90 |
| 3 | 240 | 80 |
| 4 | 150 | 95 |
| 5 | 163 | 100 |
| 6 | 154 | 95 |

Table 2. Data set so that the child is an outlier with respect to both variables- height and weight

|  |  |  |
| --- | --- | --- |
| **Child** | **Height(cms)** | **Weight(lbs)** |
| 1 | 160 | 110 |
| 2 | 158 | 90 |
| 3 | 240 | 140 |
| 4 | 150 | 95 |
| 5 | 163 | 100 |
| 6 | 154 | 95 |

Use the following stock price data (in dollars) for Exercises 13–18.

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

Q.13. Calculate the mean, median, and mode stock price.

**Ans.**Mean = (10+7+20+12+75+15+9+18+4+12+8+14)/12 = 204/12 = 17

Median = 15 & 9 = 15+9/2 = 12 (middle value) since number of points is odd then middle is chosen. if points are even the avg of two middle prices is used to determine, that is what we did here.

Mode =12 (most repeated/ high frequency)

Q, 14. Compute the standard deviation of the stock price. Interpret what this number means.

**Ans.** Standard deviation = √Σ(

= √3900/(12)

= √325

= 18.03

Standard Deviation = SD = √ σ = √325 = 18.0277563

17-18.03 =-1.03

and ($17+$18.03) = $35.03.

As we can see, every stock is priced inside this range, apart from the $75 stock.

Q.15. Find the min-max normalized stock price for the stock worth $20.

**Ans.** Min max normalized stock price for stock - $20

1. 20 12 75 15 9 18 4 12 8 14

Formula -   
X∗ = X−min(X) = X−min(X)   
mm range(X) max(X) − min(X)

Min = 4

Max = 75

Range = max-min = (75-4) = 71

**MinMaxXi=[Xi - Min(X)] / [Max(X) - Min(X)]**

Xmm = 20-4/71

= 16/71

= 0.2254  
Min-max normalized stock price for stock 20 is = **0.2254**