Assignment 2 - Theoretical Questions

Mohammad Hossein Basouli

April 6, 2025

Question 1

Correlation and **causality** are two different concepts in decission making, often confused with each other. **Correlation** only specifies whether if there is a relationship between the two variables, where **causality** talks about the effect that a cause happens to make. The **correlation** could be even meaningless, e.g. increase of *divorce* along with increase of *house price*. Also the **correlation** (if it actually is a meaningful **correlation** and not just a coincidence!) doesn't specify which of the variables causes the other or if there is a third variable in between, that is causing these two variables.

Example: assume that the number of sunburns in a town recorded for a month has a positive **correlation** with number of ice creams eaten. This obviously is a meaningful **correlation**, because we expect these two to increase together; but we can't say that number of ice creams eaten causes the number of sunburns or the other way around because the sunny weather is actually the cause of these two to increase together.

Question 2

(I couldn't label the items with alphabets so we go with numbers.)

1. **Issues of the raw data**:
    - **Data incompleteness**
    - **Data noisyness**
    - **Data inconsistency**
2. **Four major tasks in data preprocessing**:

- **Data Integration**: Gathering the data from multiple sources and integrating them.
- **Data Cleaning**: Detecting and handling missing values, outlies, noisy data, etc.
- **Data Transformation**: Aggregation and Normalization.
- **Data Reduction**: Removing unnecessary features, droping some of the rows, etc.

3. **Ways to handle missing values**:
   - **Dropping the rows**
   - **Imputation**:
     - Filling in the missing values by mean or
     - By a constant.
     - Mean of the class that the data belongs to.

Question 3

Binning could be used to smooth the noise in the data, it moves the data points that are in the same bin (or a close interval) towards the mean, median or the boundary of the bin. **Example**: Consider the data points [5, 6, 8, 100, 7, 6, 5]. We first sort the data and then put them into the bins of size 3. Here is how it would look like.
Bin 1: [5, 6, 8] $\rightarrow$ Mean = 6.33 $\rightarrow$ Smoothed bin: [6.33, 6.33, 6.33]
Bin 2: [100, 7, 6] $\rightarrow$ Mean = 37.67 $\rightarrow$ Smoothed bin: [37.67, 37.67, 37.67]

Question 4

1. **Importance of data quality in EDA and common issues such as outliers and data inconsistency**:
   - **data quality is important in following aspects**:
     - extracting accurate insights from the data analysis.
     - performance of the model.
   - **issues with outliers**:
     - could affect statistical measures badly. (data skewness, mean, standard deviation, etc.)

  - causes misinterpretations from the visulization. (misinterpretation of the distribution of the data, etc.)
- **Issues with inconsistencies**:
  - could lead to misunderstanding of the data. (Consider having different representations of a single category in our data; like having [NA, New York, na, ...] to represent a single city. They are all the same city but are represented as different values.)
  - leading to difficulty in the analysis.

2. A scenario where data quality issues lead to misleading conclusions in a real-world analysis: Assume that we have a dataset of the cars insured by an insurance company, and the dataset has been recorded by multiple different operators who don't agree to follow a similar convention while recording car models. Consider the car model *Kia Pride CD5 (1999)*, one could record it as *1999 Kia Pride*, the other one as *Pride 5 doors*, and so on. This would lead us to have many different representations of a single model of car, increasing the complexity when we want to put up car model as a feature in our model our data analysis.

3. **EDA techniques for identfying and addressing the data quality issues**:

   - **outliers issues**: with the help of visulization, smoothing, etc. we can get rid of outliers in our dataset.
   - **inconsistency issues**: database integration and management techniques could help to get rid of inconsistency issues.

Question 5

- **Min-Max Normalization**: $x' = \frac{x - \min(x)}{\max(x) - \min(x)}$ it begin the data in the range [0, 1], and also keeps the original distribution of the data.
- **Standardization**: $x' = \frac{x - \mu}{\sigma}$ this would transform a normal distribution with mean μ and std σ to z distribution.
- **Robust Scaling**: $x' = \frac{x - \text{median}(x)}{\text{IQR}(x)}$ this ia a bit more less sensitive to outliers compared to Min-Max Normalization

Question 6

The goal of **data reduction** is to reduce the data in volume while still maintaining critical information about the data. Different techniques used in this approach are as follows.

- **Data Cube Aggregation**: aggregation of multiple rows to produce a single new row, representing the other information in those rows.
- **Dimentionality Reduction**: removing highly correlated columns in order to simplify the dataset.
- **Numerosity Reduction**: using clustering to divide the dataset into multiple different clusters and then selecting only a few data points in those clusters (it could be just the mean of that cluster) to represent the other data points in the cluster.

Question 7

1. Because visulization is one of the most natural, simplest and most powerful ways to convey ideas and insight to others. An specfic concept gets undertood so much easier through visulization most of time.
2. Consider this example: Figure 1 shows an uncolorized, unmarked useless plot showing the the line plot of hire, promotion and transfers in a company on different times; It doesn't give us a whole lot of details about these variables such as do we have a special pattern in our hires over time, etc. . But Figure 2 shows a well guided plot, which essentially contains within different stories about the data, what months represent the peaks in our hires ? is there a pattern such as what months in a year are having a significantly higher hires? This plot captures the answer to all of this information. e.g. it tells us that the first month in each quarter of the year has a peak in the hires compared to the other months.

Figure 1: Traditional Visualization

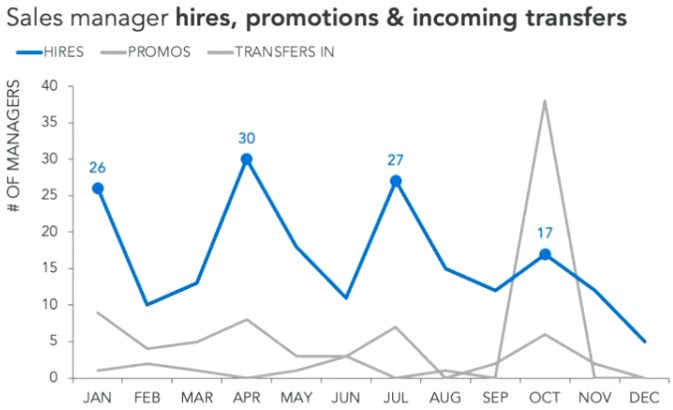

Figure 2: Visulization with Storytelling

3. **What elements transform a chart into a compelling story telling**:

   - Annonations
   - Highlights and colors
   - Well chosen and oriented labels for axis
   - Descriptive title

Question 8

1. **factors that determine the best type of chart to use for a dataset**:
    - data type of the variables. (Whether they are ordinal/nominal categorical or numerical)
    - our needs and the questions we have about the dataset. (e.g. if we want to see whether if there is a monotonic relationship between two variables, we can use scatter plot of these two variables.)
    - number of variables
    - data density and size
2. **importance of distribution charts in EDA**:
    - detecting outliers and anomalies.
    - getting insight about the original distribution of the data.
    - identifying relationship between variables.
    - helps in forming hypothesis about the data.
3. **How a heatmap of a correlation matrix can help identify patterns in a multivariate dataset:** A heatmap of the correlation matrix visually represents the pairwise correlations between variables using color intensity. More intense or darker squares typically indicate stronger positive or negative correlations between the corresponding variables. This makes it easier to detect linear relationships, clusters of related variables, or potential multicollinearity at a glance.

Question 9

pie chart

- **insight**: Compares the percentages of data in each of the different category levels.
- **suitable data types**: categorical

line chart

- **insight**: How the trend of the data might look like as the time passes.
- **suitable data types**: two variable - numerical and numerical

bar chart

- **insight**: How much data in each of the category levels resides and how the overal distribution of the variables looks like.
- **suitable data types**: one variable - categorical