# CMPE 251 Data Analytics
# Assignment 1: Data Wrangling and Analysis (Theoretical)

## 1. Handling Missing Values

A missing value is defined as data which has no recorded feature value within an instance of interest. Data can be missing because of incomplete entries, faulty equipment, corrupted files or a number of other reasons. The issue of missing data is common in any research and may have significant effect on conclusions drawn. Three methods for handling missing values in the dataset are listed below.

1. Listwise / case deletion

This method includes completely omitting the cases with missing data and continuing to analyze remaining data. This can also be known as complete case analysis and is the most common approach to handling missing values [1]. Some research suggests that this deletion introduces bias into the estimation of parameters. If the sample of data is sufficiently large in which power is not an issue listwise deletion is the best approach. However, in the case of the retail company we are not informed of how much data there is vs how many rows have missing values therefore it would be risky to use listwise deletion in the case that too much data may be removed.

2. Pairwise deletion

Pairwise deletion involves the elimination of only the feature-value pair for the instances which have missing data as opposed to the instance in full. If there is missing data elsewhere in the set, then the existing values are still used for statistical testing. Pairwise deletion includes all available data, so it preserves far more information than listwise deletion.

However, pairwise deletion presents a separate set of issues. For example, the parameters of the model will stand on different sets of data since certain information will be missing from some instances. This can produce differing statistics like sample size and standard errors [1].

3. Imputation techniques

Imputation is an overlaying technique used in machine learning to replace missing values in a dataset with reasonable estimation [1]. This is an excellent approach for when data samples are smaller to minimize the loss of information. There are two main categories of imputation with univariate and multivariate.

- **Univariate:** Involves replacing missing values with a constant like the mean, median, or mode produced from the dataset. It is easy to implement while also being computationally cheap. On the other hand, it ignores the relationships between features. With missing values that are not strictly random especially in the presence of a great inequality in the number of missing values for different variables this method may lead to inconsistent bias.
- **Multivariate:** Involves the use of regression models to predict the missing values based on the other data features. This imputation retains a great deal of data over deletion methods and avoids significantly altering the standard deviation or the shape of the distribution. Although while a regression imputation substitutes a value that is estimated by other variables no novel information is added. The sample size is increased, and the standard error is reduced.

For the case of the retail company, ideally one would implement multivariate imputation techniques. By using this approach, it is hopeful that the missing values like product price, quantity purchased, customer age, and customer gender will be replaced with accurate estimations given all the information available within the correlating row.

## 2. Addressing Inconsistent Data

To address the inconsistent data within the transaction date and product price fields it would be necessary to filter through these features and ensure standardization by changing all values into a consistent format. For transaction dates the following steps can be taken:

1. Identify all formats in use; in this case the company is experiencing inconsistencies between date formats where some transaction date data is in the form: 'MM/DD/YYYY' and the rest are represented as: 'DD-MM-YYYY'.
2. Convert to a standard format; since we are assuming the company is based out of North America, we would take the approach of switch all transaction dates into the following format: 'MM/DD/YYYY'.
3. Identify missing and invalid dates; there may be some instances where a date didn’t fit the approved format or certain values are missing. In these cases, a method to handle discrepancies will be utilized whether the feature is replaced or removed completely.

- With a consistent schema no data will be mistaken for unintended date values. Consistent date format allows for better time series analysis which can make it easier to identify trends in sales data.

A similar series of steps can be used to resolve discrepancies across the product price fields:

1. Detect currency differences and convert to a uniform currency; depending on the currency used the value will need to be converted to one unanimous currency. In this case

any currency symbol aside from a Canadian Dollar would be converted to CAD using the historical exchange rate, the affiliated currency given, and the date of transaction.

2. Scan for decimal issues; all product price fields will be standardized to two decimal places.
3. Identify any missing or invalid prices; there again may be some instances where the data does not fit the approved format or values are missing. In these cases, a method to handle discrepancies will be utilized whether the feature is replaced or removed completely.

- Standardizing the price to one currency makes it far easier to compare prices across regions and ensures the accuracy of sales statistics like average price, revenue and other related information [2].

## 3. Feature Transformation

I would propose that the customer age category gets binned according to width techniques as opposed to frequency. I would suggest that bins pertaining to youth/children and seniors take on their own width while all age groups in between would aim to follow an equal width approach. With that said, any individual age 0-17, where 18 is the legally determined age of adulthood, would fall into the youth bin. Furthermore, any individual 65 or older would be categorized as a senior and fall into the respective 65+ bin. For the remaining ages I would suggest the following widths for groupings: 18-24, 25-32, 33-39, 40-48, 49-56, 57-65. These bins are arranged such that each group is a range of ages representing people in similar life phases. This can make it easier in the future to analyze sales information which can be optimized in other domains like marketing and advertisements.

Moreover, in terms of product category, the categories should be represented at their most granular level. With all hierarchical headings included this way the data can be organized and visualized at the desired level whether it be department or alternately all the way down to product. This will again help with future analysis since there will be more information for business professionals to observe for sales/marketing purposes.

## 4. Binning

There are several challenges in decisions pertaining to data binning so it's important to ensure that strategy is considered before implementation. Binning will reduce the granularity of data, which in some cases can lead to the loss of valuable information, subtle trends within bins may be missed. Selecting the number of bins can also be a challenge, too many bins might introduce noise or start overfitting data.

In the case of a retail company the optimal binning strategy would be to separate sales data by quarters of the fiscal year. This is the most logical split from a business standpoint as most companies measure revenue and financial data using this time frame. Using a fiscal quarter

makes it easier form the company to draw comparisons against competing organizations or regarding their internal business goals. Alternately, if the company split bins month wise this might give a window into the slightly more subtle patterns within the sales information and can easily be analyzed in groups of three to observe quarterly revenue and sales.

## 5. Correlation Analysis

In terms of conducting a correlation analysis between the features product price, quantity purchased, and customer age several steps shall be followed to successfully process the data.

1. Data preparation: this involves collecting the data to ensure it is accurate and complete. Followed by collection the data shall be cleaned which includes handling missing values, ensuring standardized formats and removing any duplicates across the set.
2. Exploratory data analysis: investigate the data to understand what it can tell you. Identify the key relationships between features. From here it's possible to analyze the shape of the data and detect any outliers.
3. Correlation calculations: libraries like numpy and pandas within python can be used to calculate the correlation coefficients. Furthermore, matplotlib will make it possible to create informative data visualizations like correlation matrices and heatmaps.

Possible insights from the correlation between product price and quantity purchased could guide pricing strategies. It is likely there is a negative correlation between the price and the quantity sold, data visualizations will help business experts to price products for optimized revenue.

Insights from the correlation between customer age and product price will allow the company to strategically market their items for specified audiences with higher success. For a positive correlation between age and price it will become evident that the older customers get the more they are willing to splurge on more expensive items. With this information the company will be able to develop more accurate schemas to create advertisements tailored to different age ranges.

Finally, from the age quantity relationship the company will again gain valuable marketing insights. Where a customer in their mid 40s with a family may be likely to buy more than a 20-year-old student only supporting themselves. This information will provide more details about who is buying what and how business leaders may be able to capitalize on that.

To examine if customer location has an impact on the product categories purchased most frequently, we would again conduct a correlation analysis. This can be achieved by creating a correlation matrix to easily visualize the correlation coefficients between location and category. Furthermore, one could derive a heatmap from the matrix such that non data scientists are easily informed of the features most related and important trends.

## 6. Impact of Data Wrangling

**Handling missing values:** If missing values are improperly handled, this can immediately skew any further conclusions drawn from the data. Deleting too much data may result in a very bare dataset which doesn't properly represent the original consensus. Improperly or inaccurately replacing data instances may cause bias in the dataset and therefore further interpretation will also be inaccurate.

**Addressing inconsistent data:** This step ensures consistency and standardization across your data. If this is improperly executed one may end up with data that still has inconsistencies, and or has certain data items which have been poorly transformed and are therefore poisoning the dataset.

**Feature transformation:** Incorrect scaling and or encoding can negatively impact a dataset. It is possible that this may cause the loss of information, obscure underlying patterns in the data or even skew the training of a machine learning model.

**Binning:** Binning is a strategy used to make it easier to understand and generalize data such that relevant conclusions can be drawn from certain groups. However, if the groups themselves are poorly constructed then we simply have a pool of unrelated information where features are not as correlated as anticipated. No useful conclusions can be taken from a group of unrelated data.

**Correlation analysis:** Developing poor correlation analysis will lead us to believe certain features are either more related than they actually are or we might have two or more features highly correlated that receives a skewed analysis and gets completely missed. This can be damaging for a business because recognizing these patterns can have a high impact on the way sales, marketing and financial decisions are made in the future.

Poor data wrangling practices may manipulate the data such that incorrect conclusions are drawn. The goal of a data scientist is to generate valuable insights which can guide future business decisions. With improper wrangling techniques the information may get skewed and lead management to improperly price and market items and have an overall misguided outlook on their business scheme.

## 7. Ethical Considerations

Considering ethical implications is absolutely vital whenever handling the personal data of an individual. Data scientists have a moral responsibility to protect any personal information offered up by the general public and must take high precaution to ensure their data does not end up in the wrong hands. Security measures and access rights need to be properly managed so only the qualified personnel are handling personal data.

# Bibliography

[1] H. Kang, "The Prevention and Handling of the Missing Data," National Library of Medicine , 24 May 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3668100/. [Accessed 17 September 2024].

[2] Data Head Hunters, "Handling Inconsistent Data: Strategies for Standardization," 7 January 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3668100/. [Accessed 19 September 2024].