ALLO HEALTH ASSIGNMENT

DATA CLEANING AND PREPARATION

Dissertation submitted in fulfilment of the requirements for the Degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

By

K. Shyam

12111001

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab

DECLARATION STATEMENT

I hereby declare that the research work reported in the dissertation/dissertation proposal entitled "**ALLO HEALTH ASSIGNMENT"** in partial fulfilment of the requirement for the award of Degree for Master of Technology in Computer Science and Engineering at Lovely Professional University, Phagwara, Punjab. I have not submitted this work elsewhere for any degree or diploma.

I understand that the work presented herewith is in direct compliance with Lovely Professional University's Policy on plagiarism, intellectual property rights, and highest standards of moral and ethical conduct. Therefore, to the best of my knowledge, the content of this dissertation represents authentic and honest research effort conducted, in its entirety, by me. I am fully responsible for the contents of my dissertation work.

Signature of Candidate

K. Shyam

R.No 12111001

INTRODUCTION:

Data cleansing is the process of correcting or removing inaccurate, incomplete, incomplete, duplicate or incomplete data from a database. When combining data from multiple sources, data can easily be duplicated or mislabeled. If the data is incorrect, the results and algorithms will not be reliable, even if they appear to be correct. There is no way to write down the exact steps in the data cleaning process because the data cleaning process can vary for different datasets. But it's important to create a model for the data cleansing process so you know you're always doing it right. Data conversion is the process of converting data from one format or format to another. The transformation process, which may be called data blending or data blending, transforms and maps data from one "raw" data format to another for storage and analysis. This article focuses on the data cleansing process.

METHODOLOGY:

1. Clean the provided data sets to ensure there are no missing or inconsistent values:

Data cleaning is making sure that a given dataset doesn't have null values, removing duplicates and inconsistent data types. Simply in data cleaning it cleans those null values, removing duplicates and inconsistent data types.

Step 1: Handling the missing values:

NAN values as they can change the result during analysis machine learning models are not trained to deal with these values. This problem can be handled by using pandas library.

```
print(data_google.isnull().sum())

print('----')

print(data_facebook.isnull().sum())
```

I used these commands to find the null values of facebook and google ads datasets. After using these commands, I don't get any null values in the datasets.

Step 2: Finding the duplicates:

Duplicates always finds a way to influence the result in a way that it distorts the analysis and not accurately show the patterns and trends underlying. This problem can be handled by using pandas library.

```
print(data_google.duplicated().sum())

print(data_facebook.duplicated().sum())
```

I used these commands to find the duplicates from facebook and google ads datasets. After using these commands, I don't get any duplicates in the datasets.

Step 3: Finding the inconsistent datatypes:

Pandas are very important aspect of data preprocessing, that ensures data is in the appropriate format for analysis. Data is always messy because it was from various sources and in that some datatypes of some data values are in wrong format. For Instance, number values can come in 'float' or 'string'. Mixing up these formats leads to errors and wrong results.

data_google.info()

data_facebook.info()

I used these commands to find the inconsistent datatypes from facebook and google ads datasets. After using these commands, I don't get any inconsistent datatypes in the datasets.

Step 4: Checking the equal columns in both data sets:

```
FB_col=set(data_facebook.columns)
Gol_col=set(data_google.columns)
res=FB_col-Gol_col
res1=Gol_col-FB_col
print(res)
print(res1)
```

```
{'Lead To Call', 'Call (Conversion)'}
{'Call ', 'CAC', 'Lead to Call'}
```

Call and Lead to call columns are having the same purpose in both data sets but differs with a name.

So, we have to rename the column.

```
data_facebook.rename(columns={'Call (Conversion)':'Call','Lead To Call':'Lead to Call'},inplace=True)
```

CAC(Customer Acquisition column) was not present in the facebook data so we have to add the column to the facebook dataset.

```
data_facebook['CAC'] = data_facebook['Cost (INR)'] / data_facebook['Leads']
data_facebook.head()
```

2. Prepare the data for analysis by merging and aligning the date ranges across all campaigns and ad sets:

Step 5: Merging both datasets:

```
[ ] merged_data = pd.merge(data_google, data_facebook, on=['Date', 'Campaign Name', 'Ad Set Name'], how='outer', suffixes=('_google', '_facebook'))
    merged_data.head()
```

Merging these two datasets by combining them horizontally because combining makes it one. By horizontally means we can distinguish both datasets when they are side by side.

Step 6: Aligning the date ranges across all campaigns and all ad sets:

```
start_date = max(data_google['Date'].min(), data_facebook['Date'].min())
end_date = min(data_google['Date'].max(), data_facebook['Date'].max())

merged_data = merged_data[(merged_data['Date'] >= start_date) & (merged_data['Date'] <= end_date)]
```

CONCLUSION:

First, we identified and handled missing values using the pandas library. By running checks with data_google.isnull().sum() and data_facebook.isnull().sum(), we confirmed that there were no null values in either dataset, ensuring that our data was complete.

Next, we addressed the issue of duplicates, which can distort analysis results and obscure true patterns. Using data_google.duplicated().sum() and data_facebook.duplicated().sum(), we verified that there were no duplicate entries in the datasets.

We then examined and corrected inconsistent data types. By utilizing data_google.info() and data_facebook.info(), we ensured that all data types were consistent and correctly formatted, preventing potential errors in analysis.

In addition to these steps, we aligned the datasets structurally. We standardized column names where they represented the same data but were named differently (e.g., "Call" in Google Ads and "Lead to call" in Facebook Ads). We also added a "CAC (Customer Acquisition Cost)" column to the Facebook dataset to match the Google dataset, ensuring both datasets had a consistent structure.

Following this alignment, we merged the two datasets horizontally, combining them side-by-side. This unified dataset allows for comprehensive analysis while retaining the ability to distinguish between data sources.

Lastly, we aligned the date ranges across all campaigns and ad sets, which is essential for accurate temporal analysis and performance comparisons.

In conclusion, through meticulous data cleaning, we created an integrated and reliable dataset ready for analysis. By addressing missing values, duplicates, inconsistent data types, and structural differences, we ensured data integrity and usability, providing a solid foundation for subsequent analytical tasks.