PROG 8460

Web Analytics and Business Intel

ASSIGNMENT #1

DUE DATE: October 11th, 2021, 11:59pm

Task 1 – Data Cleaning

* **Sort Answers**
  1. **How many data records are in your final dataset?**

After cleaning process, total 11 column left in dataset.

* 1. **How many data records did you remove during the cleansing process?**

In original dataset, there were total 204 records available, after cleaning process there are total 199 records in final cleaned dataset. Total 5 records are compromised during the cleaning process.

* 1. **How many countries of origin are contained in your final dataset? List them.**

In original dataset there were 204 total records and out of total 84% missing values for column named “Country”, after applying cleaning process and missing values fill operations missing value percentage dropped to only 5% for column named “Country”.  
**List**: united states, netherlands, japan, germany, china, thailand, India, united kingdom,

italy, europe, france, romania, greece, cyprus, indonesia, mexico, costa rica, papua new guinea,egypt,iran, present-day uzbekistan, present-day france, usa

Task 2 - BI and Data related Question

1. **What are the 3 Vs of Data?**

There are 3 Vs to understand big data problems: Volume, Velocity, and Variety. These Vs are key component to understand big data problem.

1. **List capabilities of Business Intelligence systems.**

* Filtering and Transforming data

Data Filtering contains process of re-organizing, transforming and aggregating data. Transformation creates data in proper structure such as files or tables.

* Preparing Dashboards and reports

Reports are used to show large amount of data in tabular form, charts and dashboards. Dashboards used to represent KPIs Key Performance indicators using Graphical User Interface.

* Preparing and Cleaning Data

For analysing data, first data must be collected from different related sources, once data is collected BI tools helps to preparing and cleaning this unstructured data by filtering matching data and unmatching data deletion to make data appropriate for data analysis.

* Visualizing Data

Data Visualization is core of BI, where large amount of data converts into informative image which convey meaning. Visualization can be done in different type of charts: Pie, Scatter and bar.

* Dashboard Analytics

Dashboard Analytics represents large volume of data to understand trends and find out insights. This helps to find out current trends and how strategy working in business and helps to take future decision based on it.

* Using Predictive Model

Predictive modelling is creating, testing and validation model for best predictive outcome. It’s a one of core part of BI which helps organization to predict future need of staff, market demand, sales and profit/loss.

* Data Lineage Tracking

Data Lineage is process of tracking data, from its source to where data moves throughout its life cycle. It generates a complete and continuous record of system activity which can be represented visually.

1. **Different types of data with example for each type.**

* Nominal Data

E.g.: Gender, Hair Colour, Ethnicity

* Ordinal Data

E.g.:

* First, Second and Third
* A, B, C
* Low, Medium and High
* Discrete Data

E.g.: Number of students in class, number of workers in company

* Continuous Data

E.g.: Height of children, speed of cars, square feet of bedroom

1. **Define data visualization.**

Data Visualization is the process of translation huge amount of data sets and metrics into visuals forms such as graphs, charts and other visual forms, which helps to identify and share current trends and outliers and find out new insights.

1. **What is a BI system?**

Business intelligence system combines of data visualization, data mining, business analytics, data tools and infrastructure to help organization to make accurate and data driven decisions.

1. **What are the 5C's of Data for data preparation and the purpose of each?**

List of 5C’s described below with its purposes.

* Clean

Cleaning data is important part of data processing, clean data means there is no missing values, no typos, accurate data, and no outliers. Its tough task to achieve 100% clear data.

* Consistent

Consistent means all is in same manner, no matter in which database or column it is. Lets have example if Country is USA in database, then in all databases and tables Country USA should be USA only, it cant be United States or United State of America or America. It should be USA only in all places of organization.

* Conformed

Confirmed data means it should be match with column named in terms of tables. Lets take one example if we have reference of ounces mean it would be liquid, solid or Avoirdupois. It cant be other than that for ounces. It should be straightforward and universal just like year, planet names and etc.

* Current

Current data means real time data, in example stock market data it should be real time as it changes quickly.

* Comprehensive

Comprehensive in terms of data, means data should cover all dimensions of business cases. For example if we collects user data then First, last name and middle name are usually standards for names. With that sometime there is associated values such as Jr., Sr., Er. Dr., etc. Also Data should be in sufficient depth for analysing.

References

<https://www.predictiveanalyticstoday.com/key-capabilities-of-business-intelligence-software/>

<https://www.intellspot.com/data-types/>

<https://www.ibm.com/analytics/data-visualization>

<https://www.tableau.com/learn/articles/business-intelligence>

<http://kottmannconsulting.com/Talking%20Technology/five_cs_of_data.cfm>