UNIVERSITY PARTNER

UNIVERSITY OF WOLVERHAMPTON

HERALD COLLEGE KATHMANDU

Big Data
(6CS030)

Individual Coursework Report

Student Id	: 2036886
Student Name	: Sakshyat Sharma
Module Leader	: Jnaneshwar Bohara
Cohort/Batch	: 4

Submitted on : 24-04-2021

Acknowledgement

This report is an outcome of the inspiration and support of our lecturer Mr. Jnaneshwar Bohara and tutor Mr. Chiranjivi Khanal. I take this opportunity to express my profound gratitude and deep regards to my tutor for his exemplary guidance and monitoring throughout the completion of this report. I sincerely extend our hearty thanks to him, for his constant encouragement and immense motivation, which has sustained my efforts at all the stages of this report. Next, I would also like to extend our thankfulness to the University of Wolverhampton as well as Herald College for providing the module contents and good cooperation during the session.

Table of Contents

Table of Figures

1 Report

1.1 Introduction to data quality

Data quality reefers to the consistency, accuracy, uniqueness, completeness and timeliness of data in order to assess whether the data is usable for analysis. It is the measure of anomalies present in the data for determining if the data is appropriate for the task at hand. Data quality is generally determined by other factors such as its ease of manipulation, and availability. Since, data is the core factor of every organizational tasks, the quality of data plays an important role in making accurate decisions easy analysis. (Towards Data Science, 2018)

1.2 Summary of key features

1.2.1 Kim Paper

The paper explores and defines the impact of different sorts of dirty data and is impacts on data mining results. It puts forward the understanding of dirty data along with different technologies for preventing and cleaning the dirty data, as well as identifying metrics for improving the quality of data in large datasets. It focuses on the demonstration of dirty data in three different ways that are, missing data, not missing but wrong data, and not missing not wrong but unusable data. The missing data is split into two nodes, one when the null data is not allowed and other when the null data is allowed. The not-missing data node is also split into two child nodes, one when the data is wrong and unusable and other when the data is not wrong but unusable. The not-wrong but unusable data is when the same data is stored differently in more than one database. Similarly, wrong data can be distinguished into two child nodes on the basis of whether or not they can be prevented using techniques supported in relational database systems.

The different techniques for dealing with such dirty data has also been highlighted in the paper. Different commercial data quality tools can help prevent some of these dirty data types automatically, however, many of them are required to be prevented manually by the domain expert. The paper also suggests that the impact of such dirty data on data mining depends on the data mining algorithms, and also on the application that the mining process is intended for. Even a very low proportion of the dirty data might lead to wrong and inaccurate results. (KIM, et al., 2003)

1.2.2 Rahm Paper

The paper presents different data quality issues during data cleaning process present on different single-source or multi-source data collection focusing on schema-level and instance-level problems. Since data warehouses contains huge amount of data from variety of sources, it can contain many dirty data, and require better data cleaning. Data warehouse does not always work with multiple sources but single source as well. At the instance level, the problems might be spelling errors, invalid format of data, redundant values and so on. However, at the schema level problems arise when there is irrelevant schema design or when there is no suitable model-specific integrity constraints. Likewise, in multi-source, data from different sources are merged which may lead to the different outcome. At instance level, there occurs databases complications due to the naming conventions. It occurs when the properties in different databases are stored with the same named values, leading to duplicate and conflicted data.

The paper also presents different techniques for data cleaning. The phases include data analysis, defining transform and workflow rules, verification, transform and backflow of the clean data. (Rahm & Do, n.d.)

1.2.3 Conclusion

With the analysis of both the papers presented, it can be concluded that the prime idea of both the papers is to maintain data quality. Both of the research papers present the different kinds of data quality issues found in the dataset, along with the techniques and ways to eradicate such issues from the dataset. The issues, and the way to handle those issues are almost similar in both of the papers. Hence, it can be drawn that both the papers overviews the common purpose to handle the dirty data and maintain the better data quality.

1.3 Data Quality Issues

1. **Missing Values**

   Missing values is when the certain value of the data is null. Kim paper suggests that the Null data itself is not a problem, however, if the Null data is not replaced with the correct data value, the data becomes dirty, which now becomes an issue for data mining. (KIM, et al., 2003) Similarly Rahm paper demonstrates that the missing values are because of the unavailability of the values during the data entry, leaving the filed either with a dummy data or null. Whatever the situation, the missing data gives incomplete information leading to inaccuracy while data mining. The attributes with the null data becomes unusable for extracting enough information while data mining. (Rahm & Do, n.d.) The following figure from the given sample dataset shows an example of in a dataset. Here, it can be seen that many values in the column are missing.

142	Curtis	Davies	CDAVIES@example.co.uk	1/29/2005	3100	ST_CLERK		124	50
143	Randall	Matos	RMATOS@example.co.uk	15-Mar-06	2600	ST_CLERK		124	50
144	Peter	Vargas	PVARGAS@example.co.uk	9-Jul-06	2500	ST_CLERK		124	50
145	John	Russell	JRUSSEL@example.co.uk	1-Oct-04	14000	SA_MAN	0.4	100	80
146	Karen	Partners	KPARTNER@example.co.uk	5-Jan-05	13500	SA_MAN	0.3	100	80
147	Alberto	Errazuriz	AERRAZUR@example.co.uk	10-Mar-05	12000	SA_MAN	0.3	100	80
148	Gerald	Cambrault	GCAMBRAU@example.co.uk	15-Oct-07	11000	SA_MAN	0.3	100	80
149	Eleni	Zlotkey	EZLOTKEY@example.co.uk	29-Jan-08	10500	SA_MAN	0.2	100	80
150	Curtis	Davis	CDAVIES@example.co.uk	29-Jan-05	3100	ST_CLERK		124	50
151	David	Bernstein	DBERNSTE@example.co.uk	24-Mar-05	9500	SA_REP	0.25	145	80

Figure 1 Missing Values

2. Duplicated records

Duplicated records occurs when the values for the same entity are represented more than once in the table. This creates data redundancy, causing slowdowns in data analysis processes. The paper of Rahm shows that such duplicated data occurs due to some error in the data entry when same data is repeated multiple times. However, if the same entity is represented by the different values, it makes the entries contradicting. (Rahm & Do, n.d.) Likewise, Kim paper focuses on usability of such duplicate data. These kind of duplicate data might not be wrong, but are unusable. (KIM, et al., 2003) Below is an example of the duplicated records in the sample data provided. Here, as it can be seen the same record has been repeated twice in the dataset.

141	Trenna	Rajs	TRAJS@example.co.uk	17-Oct-03	3500	ST_CLERK		124	50
142	Curtis	Davies	CDAVIES@example.co.uk	1/29/2005	3100	ST_CLERK		124	50
143	Randall	Matos	RMATOS@example.co.uk	15-Mar-06	2600	ST_CLERK		124	50
144	Peter	Vargas	PVARGAS@example.co.uk	9-Jul-06	2500	ST_CLERK		124	50
145	John	Russell	JRUSSEL@example.co.uk	1-Oct-04	14000	SA_MAN	0.4	100	80
146	Karen	Partners	KPARTNER@example.co.uk	5-Jan-05	13500	SA_MAN	0.3	100	80
147	Alberto	Errazuriz	AERRAZUR@example.co.uk	10-Mar-05	12000	SA_MAN	0.3	100	80
148	Gerald	Cambrault	GCAMBRAU@example.co.uk	15-Oct-07	11000	SA_MAN	0.3	100	80
149	Eleni	Zlotkey	EZLOTKEY@example.co.uk	29-Jan-08	10500	SA_MAN	0.2	100	80
150	Curtis	Davis	CDAVIES@example.co.uk	29-Jan-05	3100	ST_CLERK		124	50
151	David	Bernstein	DBERNSTE@example.co.uk	24-Mar-05	9500	SA_REP	0.25	145	80
152	Peter	Hall	PHALL@example.co.uk	20-Aug-05	9000	SA_REP	0.25	145	80

Figure 2 Duplicated Records

3. **Ambiguous Data**

Most of the data are ambiguous either due to the use of abbreviations or because of the incomplete context. According to the Kim paper, the ambiguity in data are cause by the use of abbreviations and the incomplete contexts. There might be lots of words which denotes the same abbreviations causing the ambiguity in data. Similarly, lots of the values in the data are incomplete which does not provide any contextual meaning leading to data ambiguity. (KIM, et al., 2003) Also, the paper by Rahm highlights the use of different cryptic values and abbreviations leading to ambiguous data. Analysis on such data might not give meaningful and contextual results, hence making it one of the major data quality issues. (Rahm & Do, n.d.) For example, there might occur the chances of representing two different names 'John Smith' and 'James Smith' both in the abbreviated form 'J. Smith'. Here, both the names are different but are represented in a same way. This causes data ambiguity.

2 Sample Data

The two CSV dataset employee and department has been provided as the sample dataset for analyzing the dirty data. Department dataset seems to be all fine, however, the following are the major data quality issues that have been found in the employee dataset.

1. Missing Data / Null Values

There occurs a lot of missing data in the dataset. These kind of missing data can cause bias in the estimation of parameters and inaccurate statistical results for analysis, making it one of the most common issues in data. In the given employee dataset, majority of the rows in the 'Commission_pct' column is missing. Similarly, some of the rows have missing 'dept_id' and 'manager_id' with null value, as seen as follows. Such columns having higher number of null attributes cannot be assumed by any means, hence it is removed completely.

EMP_ID	FIRST_NAME	LAST_NAME	EMAIL	HIREDATE	SALARY	JOB_ID	COMMISSION_PCT	MANAGER_ID	DEPT_ID
100	Steven	King	SKING@example.co.uk	17-Jun-03	24000	AD_PRES			90
101	Neena	Kochhar	NKOCHHAR@example.co.uk	21-Sep-05	17000	AD_VP		100	95
102	Lex	DeHaan	LDEHAAN@example.co.uk	13-Jan-01	17000	AD_VP		100	90
103	Alexander	Hunold	AHUNOLD@example.co.uk	3-Jan-06	9000	IT_PROG		102	60
104	Bruce	Ernst	BERNST@example.co.uk	21-May-07	6000	IT_PROG		103	60
105	David	Austin	DAUSTIN@example.co.uk	25-Jun-05	4800	IT_PROG		103	60
106	Valli	Pataballa	VPATABAL@example.co.uk	5-Feb-66	4800	IT_PROG		103	60
107	Diana	Lorentz	DLORENTZ@example.co.uk	7-Feb-07	4200	IT_PROG		103	60
108	Nancy	Greenberg	NGREENBE@example.co.uk	17-Aug-02	12000	FI_MGR		101	100
109	Daniel	Faviet	DFAVIET@example.co.uk	16-Aug-02	9000	FI_ACCOUNT		108	100
110	John	Chen	JCHEN@example.co.uk	28-Sep-05	8200	FI_ACCOUNT		108	100
111	Ismael	Sciarra	ISCIARRA@example.co.uk	30-Sep-05	7700	FI_ACCOUNT		108	100
112	JoseManuel	Urman	JMURMAN@example.co.uk	7/3/2006	7800	FI_ACCOUNT		108	100
113	Luis	Popp	LPOPP@example.co.uk	7-Dec-07	6900	FI_ACCOUNT		108	100
114	Den	Raphaely	DRAPHEAL@example.co.uk	7-Dec-02	11000	PU_MAN		100	30
115	Alexander	Khoo	AKHOO@example.co.uk	18-MAI-2003	3100	PU_CLERK		114	30
116	Shelli	Baida	SBAIDA@example.co.uk	24-Dec-05	2900	PU_CLERK		114	30
117	Sigal	Tobias	STOBIAS@example.co.uk	24-Jul-05	128000	PU_CLERK		114	30
118	Guy	Himuro	GHIMURO@example.co.uk	15-Nov-06	2600	PU_CLERK		114	30
119	Karen	Colmenares	KCOLMENA@example.co.uk	10-Aug-07	2500	PU_CLERK		114	30
120	Matthew	Weiss	MWEISS@example.co.uk	18-Jul-04	8000	ST_MAN		100	50
121	Adam	Fripp	AFRIPP@example.co.uk	31-APR-2005	8200	ST_MAN		100	50

Figure 3 Sample Data - Missing/Null Values

2. Invalid Data Format

The data in ‘Hiredate’ column of the given dataset has inappropriate and irregular format of data. The dates are in no particular format. Different formats of dates has been used for different rows in the dataset. This causes the problem while importing the dataset and performing queries. So the all dates should be shortened into the same format using oracle date formation. As seen in the following figure, many data in this field is not compatible with the standard date format which giver error during the import of data.

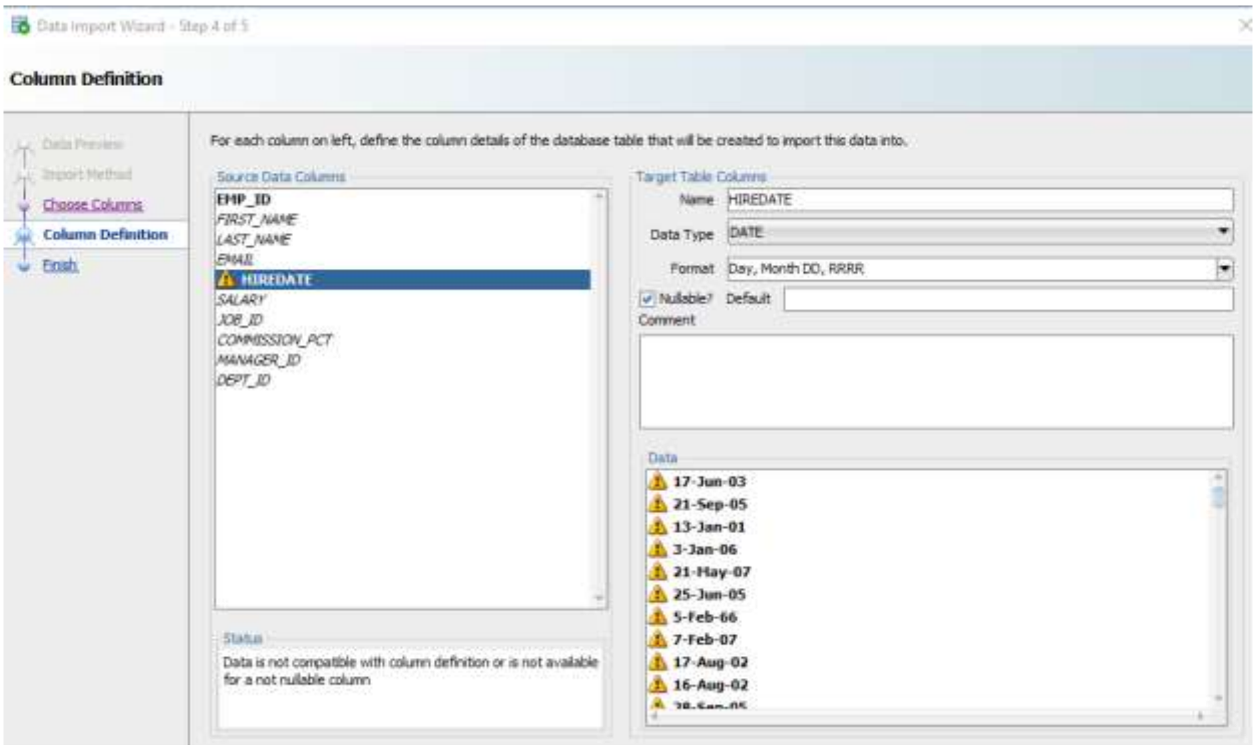

Figure 4 Sample Data - Invalid Date format

3. Data Redundancy / Duplicate Data

The repetition of same data again and again cause the duplication of data. This might lead to slower analysis of data while data mining, as the same values are repeated again and again causing redundancy. It might also create system slowdowns as same data are stored multiple times. The given sample dataset employee also has some of the repeated values in the data as shown below. For the removal for such redundancies, the data could be manually removed for the smaller tables as given. However, for the larger datasets, automated database cleanup could be used.

Here, as we can see, the employees with the exact same values has been recorded twice, which denotes the duplication of data.

3	141	Trenna	Rajs	TRAJS@example.co.uk	17-Oct-03	3500	ST_CLERK		124	50
4	142	Curtis	Davies	CDAVIES@example.co.uk	1/29/2005	3100	ST_CLERK		124	50
5	143	Randall	Matos	RMATOS@example.co.uk	15-Mar-06	2600	ST_CLERK		124	50
5	144	Peter	Vargas	PVARGAS@example.co.uk	9-Jul-06	2500	ST_CLERK		124	50
7	145	John	Russell	JRUSSEL@example.co.uk	1-Oct-04	14000	SA_MAN	0.4	100	80
3	146	Karen	Partners	KPARTNER@example.co.uk	5-Jan-05	13500	SA_MAN	0.3	100	80
9	147	Alberto	Errazuriz	AERRAZUR@example.co.uk	10-Mar-05	12000	SA_MAN	0.3	100	80
0	148	Gerald	Cambrault	GCAMBRAU@example.co.uk	15-Oct-07	11000	SA_MAN	0.3	100	80
1	149	Eleni	Zlotkey	EZLOTKEY@example.co.uk	29-Jan-08	10500	SA_MAN	0.2	100	80
2	150	Curtis	Davis	CDAVIES@example.co.uk	29-Jan-05	3100	ST_CLERK		124	50
3	151	David	Bernstein	DBERNSTE@example.co.uk	24-Mar-05	9500	SA_REP	0.25	145	80
4	152	Peter	Hall	PHALL@example.co.uk	20-Aug-05	9000	SA_REP	0.25	145	80

103	Alexander	Hunold	AHUNOLD@example.co.uk	3-Jan-06	9000	IT_PROG		102	60
104	Bruce	Ernst	BERNST@example.co.uk	21-May-07	6000	IT_PROG		103	60
105	David	Austin	DAUSTIN@example.co.uk	25-Jun-05	4800	IT_PROG		103	60
106	Valli	Pataballa	VPATABAL@example.co.uk	5-Feb-66	4800	IT_PROG		103	60
107	Diana	Lorentz	DLORENTZ@example.co.uk	7-Feb-07	4200	IT_PROG		103	60
108	Nancy	Greenber	NGREENBE@example.co.uk	17-Aug-02	12000	FI_MGR		101	100
109	Daniel	Faviet	DFAVIET@example.co.uk	16-Aug-02	9000	FI_ACCOUNT		108	100
197	Kevin	Feeney	KFEENEY@example.co.uk	23-May-06	3000	SH_CLERK		124	50
198	Donald	OConnell	DOCONNEL@example.co.uk	21-Jun-07	2600	SH_CLERK		124	50
199	Douglas	Grant	DGRANT@example.co.uk	13-Jan-08	2600	SH_CLERK		124	50
200	David	Austen	DAUSTIN@example.co.uk	25-Jun-05	4800	IT_PROG		103	60
201	Michael	Hartstein	MHARTSTE@example.co.uk	17-Feb-04	13000	MK_MAN		100	20
202	Pat	Fay	PFAY@example.co.uk	17-Aug-05	6000	MK_REP		201	20
203	Susan	Mavris	SMAVRIS@example.co.uk	7-Jun-02	6500	HR_REP		101	40

Figure 5 Sample Data - Duplicate Records

4. Misspelled Data

Some of the data values are misspelled in the given employee dataset. This might cause difficulties in analysis because even the data with the same values might be seen as different data due to misspelled values. Such misspelled values could be manually replaced by the correct spellings if the dataset is smaller as the given sample. Below shows some of the misspelled data that is found in the sample data.

Here, in the date column, the month 'May' has been misspelled as 'Mai'.

10	John	Chen	JCHEN@example.co.uk	28-Sep-05	8200	FI_ACCOUN
11	Ismael	Sciarra	ISCIARRA@example.co.uk	30-Sep-05	7700	FI_ACCOUN
12	JoseManuel	Urman	JMURMAN@example.co.uk	7/3/2006	7800	FI_ACCOUN
13	Luis	Popp	LPOPP@example.co.uk	7-Dec-07	6900	FI_ACCOUN
14	Den	Raphaely	DRAPHEAL@example.co.uk	7-Dec-02	11000	PU_MAN
15	Alexander	Khoo	AKHOO@example.co.uk	18-MAI-2003	3100	PU_CLERK
16	Shelli	Baida	SBAIDA@example.co.uk	24-Dec-05	2900	PU_CLERK

Figure 6 Sample Data - Misspelled Date

Here, both of the highlighted data are same data with same values, however, the last name of both of these values have different spellings one 'Davis' and other 'Davies'. So, it can be analyzed that one of them is misspelled. By looking at ta email of both the data, the correct spelling can be assumed to be 'Davies' and replaced accordingly.

141	Trenna	Rajs	TRAJS@example.co.uk	17-Oct-03	3500	ST_CLERK		124	50
142	Curtis	Davies	CDAVIES@example.co.uk	1/29/2005	3100	ST_CLERK		124	50
143	Randall	Matos	RMATOS@example.co.uk	15-Mar-06	2600	ST_CLERK		124	50
144	Peter	Vargas	PVARGAS@example.co.uk	9-Jul-06	2500	ST_CLERK		124	50
145	John	Russell	JRUSSEL@example.co.uk	1-Oct-04	14000	SA_MAN	0.4	100	80
146	Karen	Partners	KPARTNER@example.co.uk	5-Jan-05	13500	SA_MAN	0.3	100	80
147	Alberto	Errazuriz	AERRAZUR@example.co.uk	10-Mar-05	12000	SA_MAN	0.3	100	80
148	Gerald	Cambrault	GCAMBRAU@example.co.uk	15-Oct-07	11000	SA_MAN	0.3	100	80
149	Eleni	Zlotkey	EZLOTKEY@example.co.uk	29-Jan-08	10500	SA_MAN	0.2	100	80
150	Curtis	Davis	CDAVIES@example.co.uk	29-Jan-05	3100	ST_CLERK		124	50
151	David	Bernstein	DBERNSTE@example.co.uk	24-Mar-05	9500	SA_REP	0.25	145	80
152	Peter	Hall	PHALL@example.co.uk	20-Aug-05	9000	SA_REP	0.25	145	80

Figure 7 Sample data - Misspelled last name 'Davies'

Similarly, the following two highlighted data are same as well but the last names of both data are spelled differently one ‘Austin’ and the other one ‘Austen’. Here, one of them is misspelled. By looking at the email of both of these data, the correct spelling is assumed to be ‘Austin’, and replaced accordingly.

103	Alexander	Hunold	AHUNOLD@example.co.uk	3-Jan-06	9000	IT_PROG		102	60
104	Bruce	Ernst	BERNST@example.co.uk	21-May-07	6000	IT_PROG		103	60
105	David	Austin	DAUSTIN@example.co.uk	38528	4800	IT_PROG		103	60
106	Valli	Pataballa	VPATABAL@example.co.uk	5-Feb-66	4800	IT_PROG		103	60
198	Donald	OConnell	DOCONNEL@example.co.uk	21-Jun-07	2600	SH_CLERK		124	50
199	Douglas	Grant	DGRANT@example.co.uk	13-Jan-08	2600	SH_CLERK		124	50
200	David	Austen	DAUSTIN@example.co.uk	38528	4800	IT_PROG		103	60
201	Michael	Hartstein	MHARTSTE@example.co.uk	17-Feb-04	13000	MK_MAN		100	20
202	Pat	Fay	PFAY@example.co.uk	17-Aug-05	6000	MK_REP		201	20
203	Susan	Mavris	SMAVRIS@example.co.uk	7-Jun-02	6500	HR_REP		101	40

Figure 8 Sample Data - Misspelled last name 'Austin'

5. Different representation for same data

In the provided sample dataset employee, one of the records is represented with different name for the same value. As it can be seen in the figure below, the name ‘SA_REP’ and ‘SALES_REP’ both represent the same department i.e. Sales. However, the names are different in one of the data. This is also one of the major issues in dataset causing ambiguity in dataset. This can be fixed manually by replacing the name ‘SALES_REP’ as ‘SA_REP’ to make the same representation.

163	Danielle	Greene	DGREENE@example.co.uk	19-Mar-07	9500	SA_REP	0.15	147	80
164	Mattea	Marvins	MMARVINS@example.co.uk	24-Jan-08	7200	SA_REP	0.1	147	80
165	David	Lee	DLEE@example.co.uk	23-Feb-08	6800	SA_REP	0.1	147	80
166	Sundar	Ande	SANDE@example.co.uk	24-Mar-08	6400	SA_REP	0.1	147	80
167	Amit	Banda	ABANDA@example.co.uk	21-Apr-08	6200	SA_REP	0.1	147	80
168	Lisa	Ozer	LOZER@example.co.uk	11-Mar-05	11500	SALES_REP	0.25	148	80
169	Harrison	Bloom	HBLOOM@example.co.uk	23-Mar-06	10000	SA_REP	0.2	148	80
170	Tayler	Fox	TFOX@example.co.uk	24-Jan-06	9600	SA_REP	0.2	148	80
171	William	Smith	WSMITH@example.co.uk	23-Feb-07	7400	SA_REP	0.15	148	80
172	Elizabeth	Bates	EBATES@example.co.uk	24-Mar-07	7300	SA_REP	0.15	148	80
173	Sundita	Kumar	SKUMAR@example.co.uk	21-Apr-08	6100	SA_REP	0.1	148	80
174	Ellen	Abel	EABEL@example.co.uk	11-May-04	11000	SA_REP	0.3	149	80
175	Alyssa	Hutton	AHUTTON@example.co.uk	19-Mar-05	8800	SA_REP	0.25	149	80
176	Jonathon	Taylor	JTAYLOR@example.co.uk	24-Mar-06	8600	SA_REP	0.2	149	80
177	Jack	Livingston	JLIVINGS@example.co.uk	23-Apr-06	8400	SA_REP	0.2	149	80

Figure 9 Sample Data - Different Representation of same data

3 Evidence

3.1 SQL Statements and Results

Employee Dataset

Importing the data

First the employee.csv file is imported into the database as shown below:

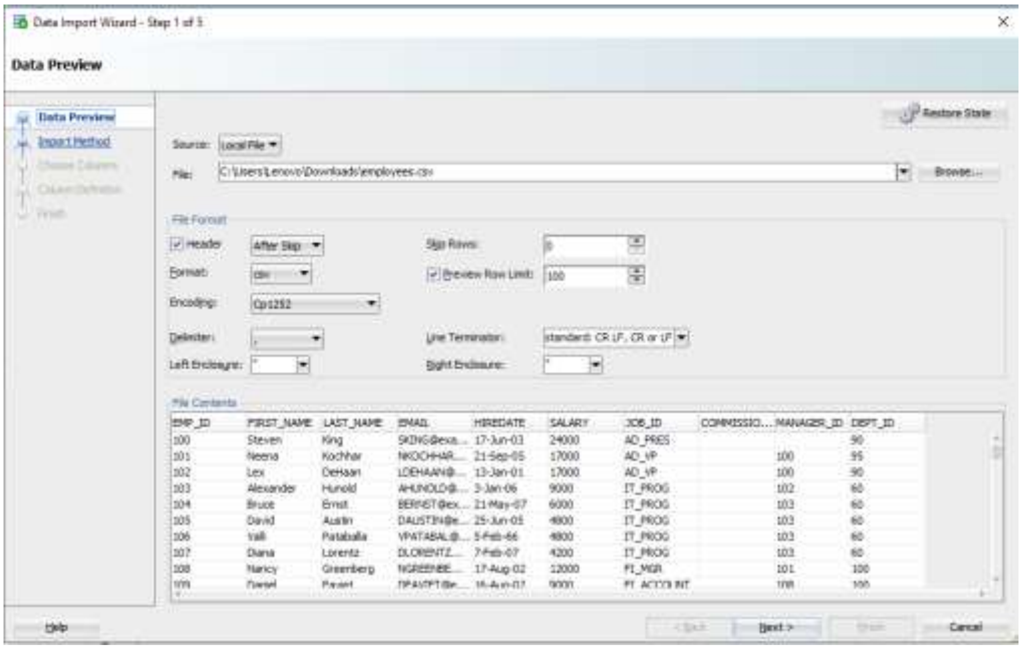

Figure 10 Evidence - Importing the 'employee' dataset

SQL Statements

The following query checks the data having the dept_id null. It returns all the data where the value in dept_id is missing.

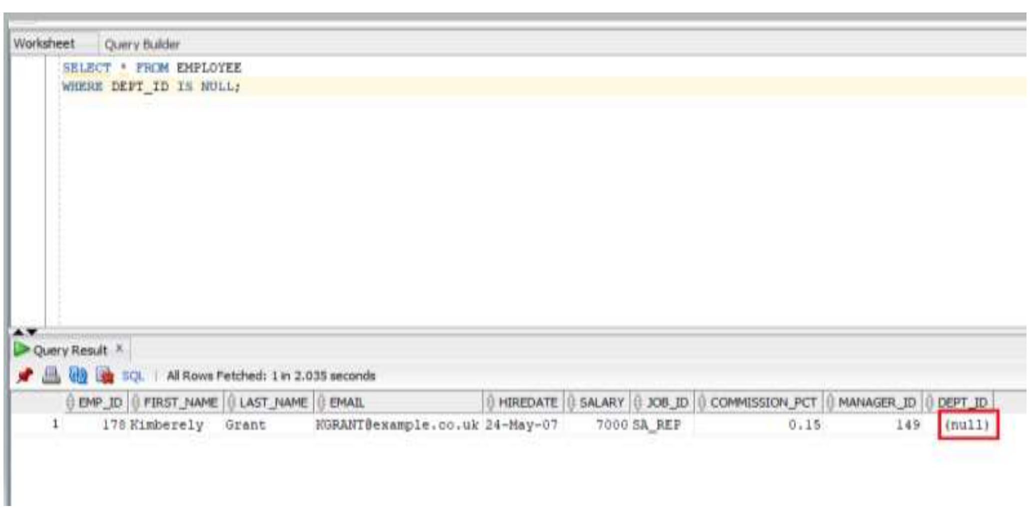

Figure 11 Evidence - SQL query to find null data in 'Dept_id'

The following query checks the data having the manager_id null. It returns all the data where the value in manager_id is missing.

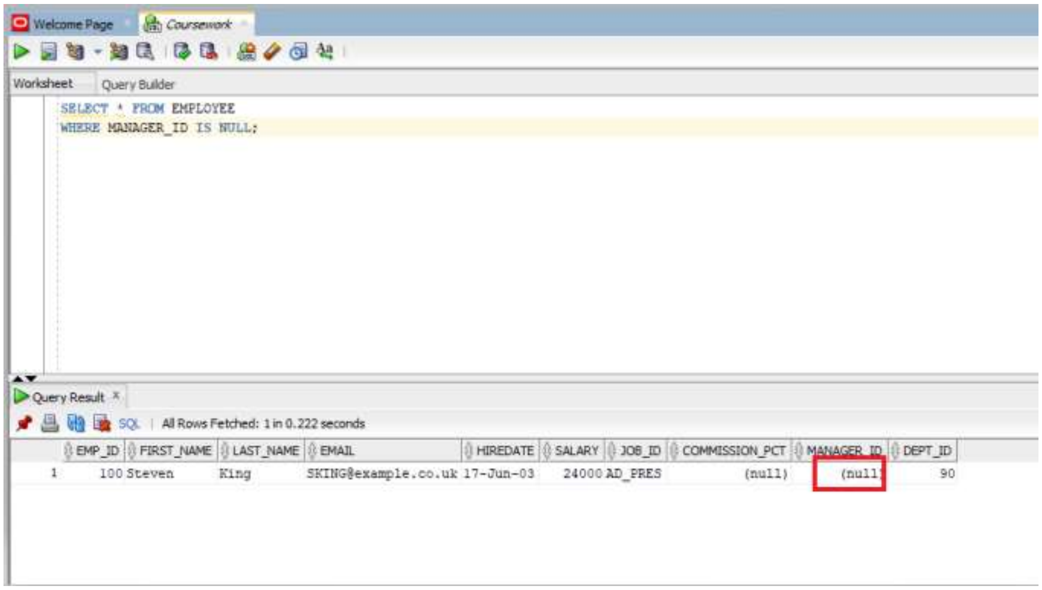

Figure 12 Evidence - SQL query to find null data in 'Manager_id'

The following query checks the data having the commission_pct null. It returns all the data where the value in commission_pct is missing.

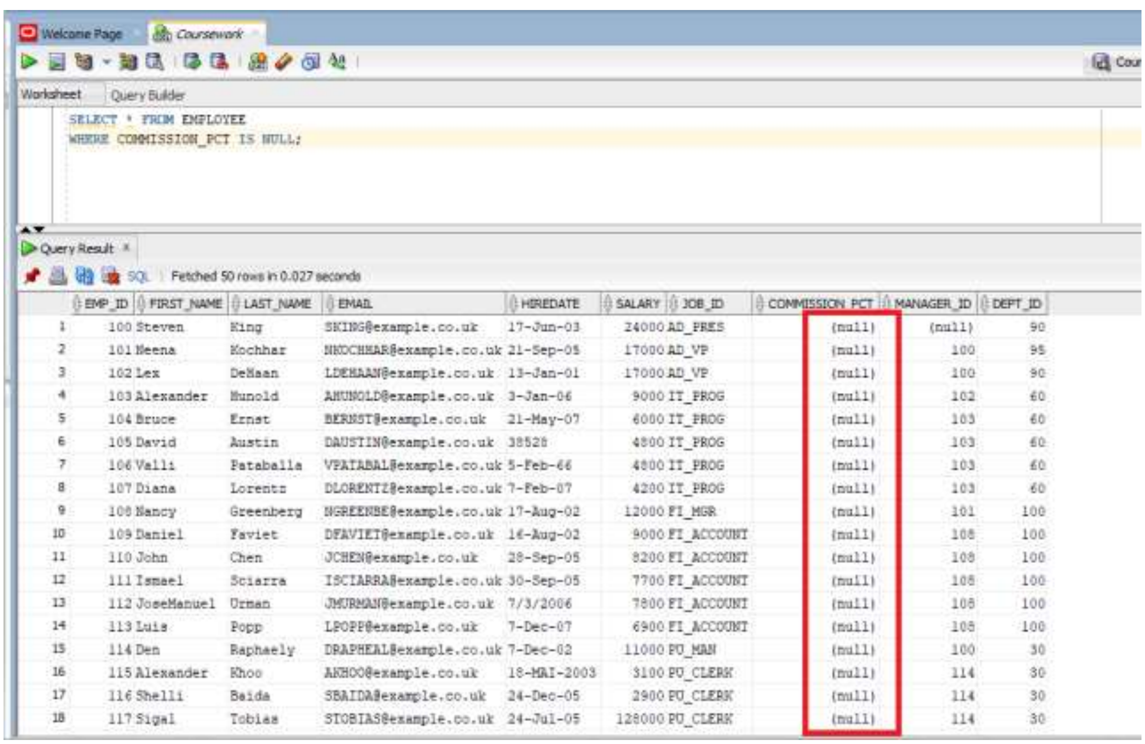

Figure 13 Evidence - SQL query to find null data in 'Commission_Pct'

The query below checks for the duplicate data entries. Here, as we know email cannot be same for multiple people. Hence, the query checks how many times the email has been repeated in order to find out the duplicate entries of data.

Here, rows with the email shown in the results are the duplicate data with same values.

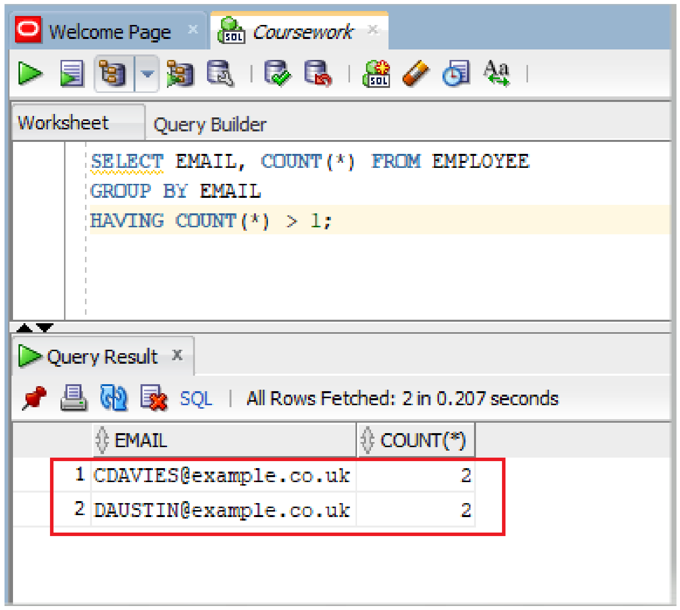

Figure 14 Evidence - SQL query to check the duplication of data using email

3.2 Visualization

Here, as it can be seen in the line graph below, it can be analyzed that lots of the values in the commission_pct column is NULL causing the line graph to collapse to zero. Only some of the fields in the column are given with values. Moreover, the difference in the value is also irregular making the data unclean and unsuitable of analysis.

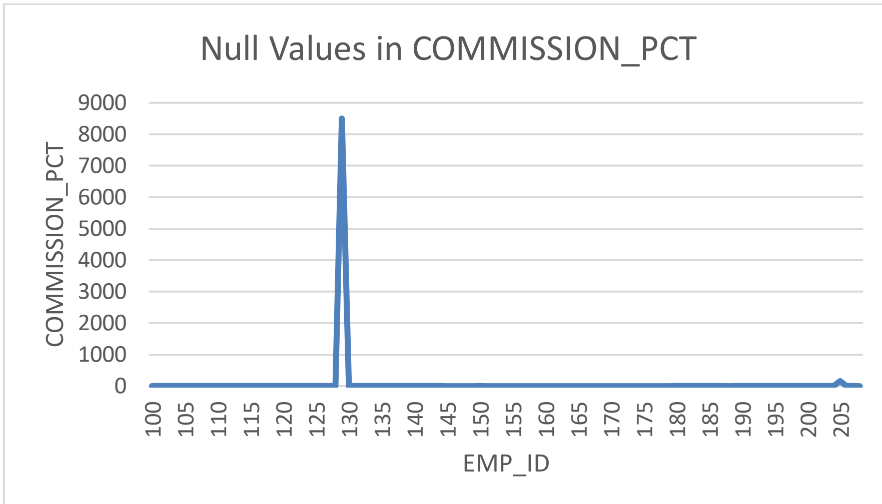

Figure 15 Visualization - Null values in 'Commission_PCT'

Similarly, the line chart below shows the presence of NULL data in the 'Manager_ID' column. Here, it can be seen that one of the data has missing values in the 'Manager_ID' column.

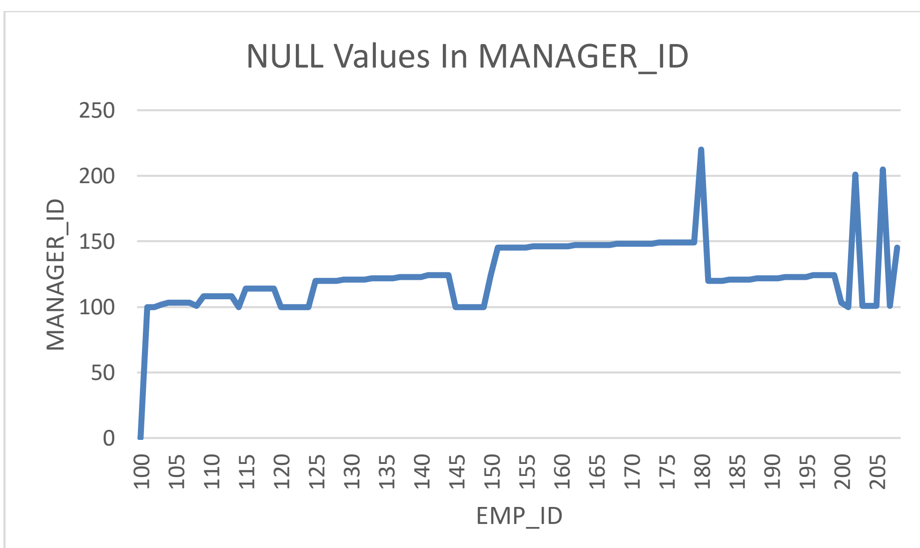

Figure 16 Visualization - Null values in 'Manager_ID'

In the same way, as it can be seen in the graph below, one of the employees has a value missing in ‘DEPT_ID’ column.

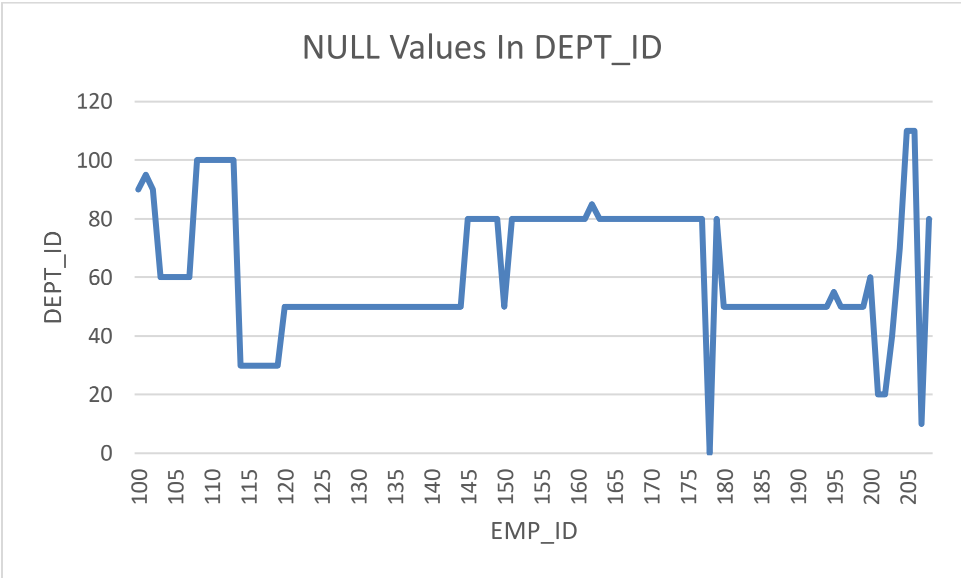

Figure 17 Visualization - Null values in 'Dept_ID'

4 References

- KIM, W. et al., 2003. A Taxonomy of Dirty Data. *Data Mining and Knowledge Discovery*, 7(1), pp. 81-99.
- Rahm, E. & Do, H. H., n.d. ‘Data cleaning: problems and current approaches. *IEEE Data Eng. Bull.*, 23(4), pp. 3-13.
- Towards Data Science, 2018. *An introduction to Data Quality*. [Online] Available at: https://towardsdatascience.com/an-introduction-to-data-quality-951cc6fe0274
  [Accessed 24 April 2021].