**Big Data**

**6CS030**

**Individual Coursework**

Student Id : 2038578

Student Name : Sandesh Thapa Magar

Group :L6CG6

Submitted to : Janeshwar Bohara

Submitted on :4/26/2021

Table of Contents

[1. Introduction to data quality 1](#_Toc70335628)

[1.1 Kim Paper 1](#_Toc70335629)

[1.2 Rahm paper 1](#_Toc70335630)

[1.3 Conclusion 2](#_Toc70335631)

[1.4 Three types of data quality issue 2](#_Toc70335632)

[2. Sample data: 3](#_Toc70335633)

[2.1 Problems: 3](#_Toc70335634)

[2.2 Solutions 4](#_Toc70335635)

[3. Evidence 5](#_Toc70335636)

[Evidence to show the issue 5](#_Toc70335637)

[Cleaning Data 8](#_Toc70335638)

[Importing 14](#_Toc70335639)

[SQL queries 15](#_Toc70335640)

[Visualization 18](#_Toc70335641)

[Bibliography 21](#_Toc70335642)

# 1. Introduction to data quality

Quality of data is the estimation of the condition of information reliant upon contemplations like exactness, trustworthiness, congruity, straightforwardness and exceptional information. Information quality measures can empower associations to identify information mistakes to be remedied and to decide whether the information in their IT measures is appropriate for the capacity it looks for (Techopedia, 2020).

## 1.1 Kim Paper

It says that data consistency has long been discussed. In this article, dirty data are essentially unfinished or incorrect data which show the same data. In order to remove or patch dirty data, data can be cleaned up earlier data processing applications. No metadata defining heritage data sources is also available here. According to this research paper, there is no extensive systematic taxonomy of dirty records. Without those taxonomic and metrics, it will always be difficult for market cleverness excellence emanating from data storehouses and the consistency of choices dependent on intellect.

The scientific classification offers an understanding into roots a full assortment of grimy information and the impact on information mining and sheds of filthy information center around messy information the executives’ strategies and evaluation measurements nature of information. We anticipate that such a taxonomy should convey a helpful guide for additional advancement Monetary examination and upgrade. This paper anticipates an important guide to additional investigation and advancement of business products in such a scientific categorization.

This exploration paper proposes several industrial big data tools offer several options for data warehousing generation and information handling for multidimensional investigation and information mining. When all is said done, clients may change missed information in a medium, medium and medium-range area (Kim, et al., 2003).

## 1.2 Rahm paper

The point of information tidy up is to distinguish and annihilate mistakes and to build information quality in debates. Issues of information consistency are found in single assortments, like chronicles and data sets. This article manages the inquiries of grouping of information substance communicated by the refinement of information and sums up the primary arrangements.

Data Cleaning problems:

The point of information tidy up is to distinguish and annihilate mistakes and to build information quality in debates. Issues of information consistency are found in single assortments, like chronicles and data sets. This article manages the inquiries of grouping of information substance communicated by the refinement of information and sums up the primary arrangements.

Single Source problem:

Having satisfactory information esteems relies to a great extent upon how much the technique and reasonableness limitations rule. There are not many limitations on information assortment and capacity for sources without plans like logs that are bound to bring about slip-ups and confusion.

Multi-source problems:

Single source issues are intensified where various sources are needed to be blended. Messy information ought to be utilized in the reason and source information can be seen, overlaid or in any case negated. The truth of the matter is that triggers are ordinarily evolved, utilized and oversaw independently to fulfill unequivocal necessities

## 1.3 Conclusion

As far as information consistency and unsanitary information, Kim and Rahm are indistinguishable high-level archives. In any case, the naming of different information issues contrasts. Kim paper discusses the scientific categorization of filthy information, which shows the effect of messy information on the consequence of information mining. Rahms' paper classified information addresses dependent on information sources, alongside Kim's paper's progression of information concerns.

## 1.4 Three types of data quality issue

* Inconsistent formats:

Numerous frameworks which battle to perceive thing in a similar classification when entering information which secures a similar substance, however which are put away in different configurations, to deliver mistaken outcomes.

* Duplicated Data

Any company must deal with this issue. If these references are mixed with the processing, multiple versions of the same data might be considerably unreliable or corrupt.

* Incomplete Information

Incomplete or generally void territories might be a vital test for assets, for example, the computerized markets and impacts industry of information-based organizations.

# 2. Sample data:

The data collection contains filthy data in the study. Data replication, incomplete data, data inconsistency, undesirable data, outdated information, and error created by human beings are problems with data sets.

## 2.1 Problems:

* Outdated/obsolete information:

Subsequently, obsolete information just alludes to old information where information isn't adjusted and updated for quite a while. Obsolete information can be troublesome and cause issues a large part of the time.

* Missing Values:

There is missing information on some columns like:

* Department\_ID
* Commission\_PCT
* Manager\_ID

It causes different problems in data analyzes, statistical power, and decreases the sample representativity. That brings one to the wrong or incorrect conclusion.

* Incorrect Data:

The sample information incorporates incorrect data, for example, Division ID 95 is given in the worker dataset yet in the department dataset. It is one to numerous associations between two arrangements of information. When fabricating the information assortment, representative and office id information is absent

## 2.2 Solutions

* Inconsistent Data:

We may make proper restrictions on the data for this issue. This issue can be resolved in the same way as DATE constraint on the date column of the same date format.

* Missing Information:

In the case of missing information, few basically erasing it will tackle the issue. Be that as it may, assuming the missing information is in huge number, the conceivable arrangement can be Recuperating the worth, Normal ascription and Various attribution.

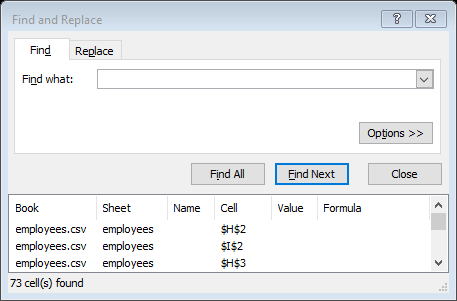
* Incorrect Data:

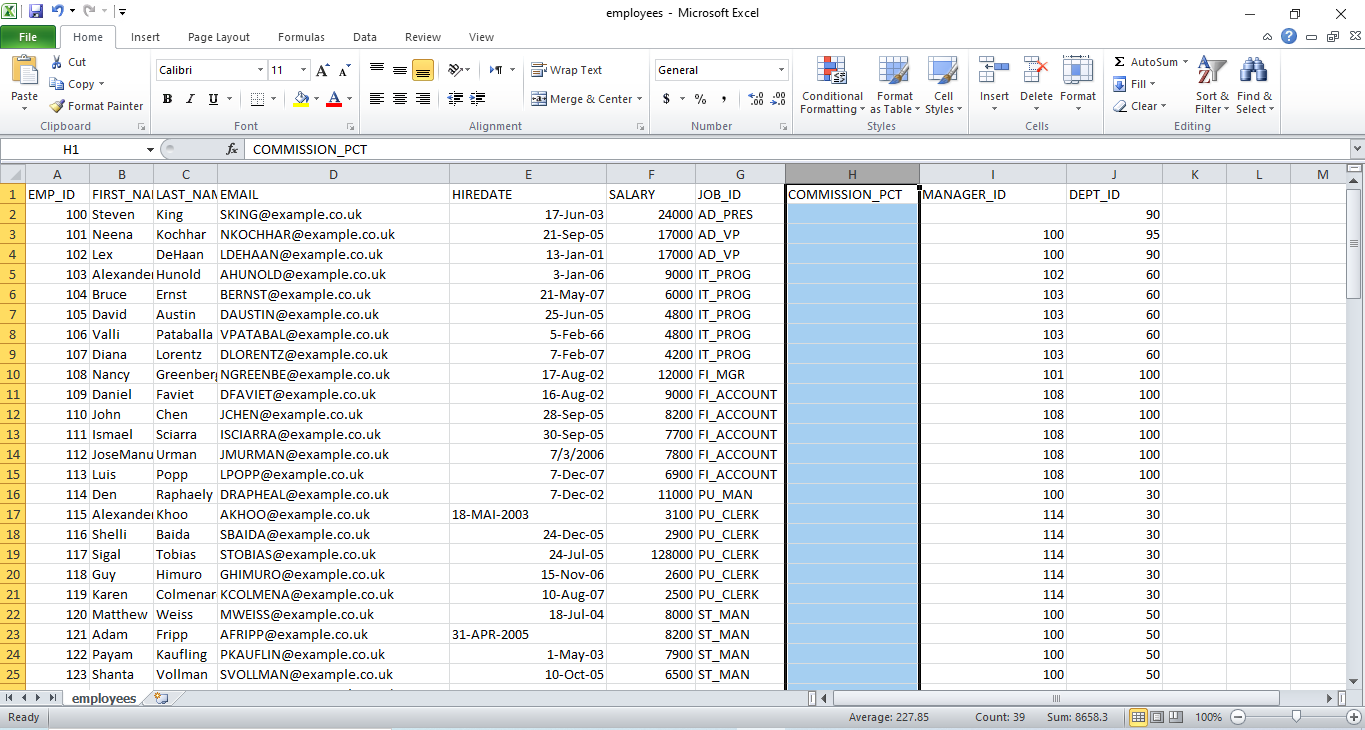
We can investigation the information and delete the wrong information and address the difficulty, if the information is in little numbers.

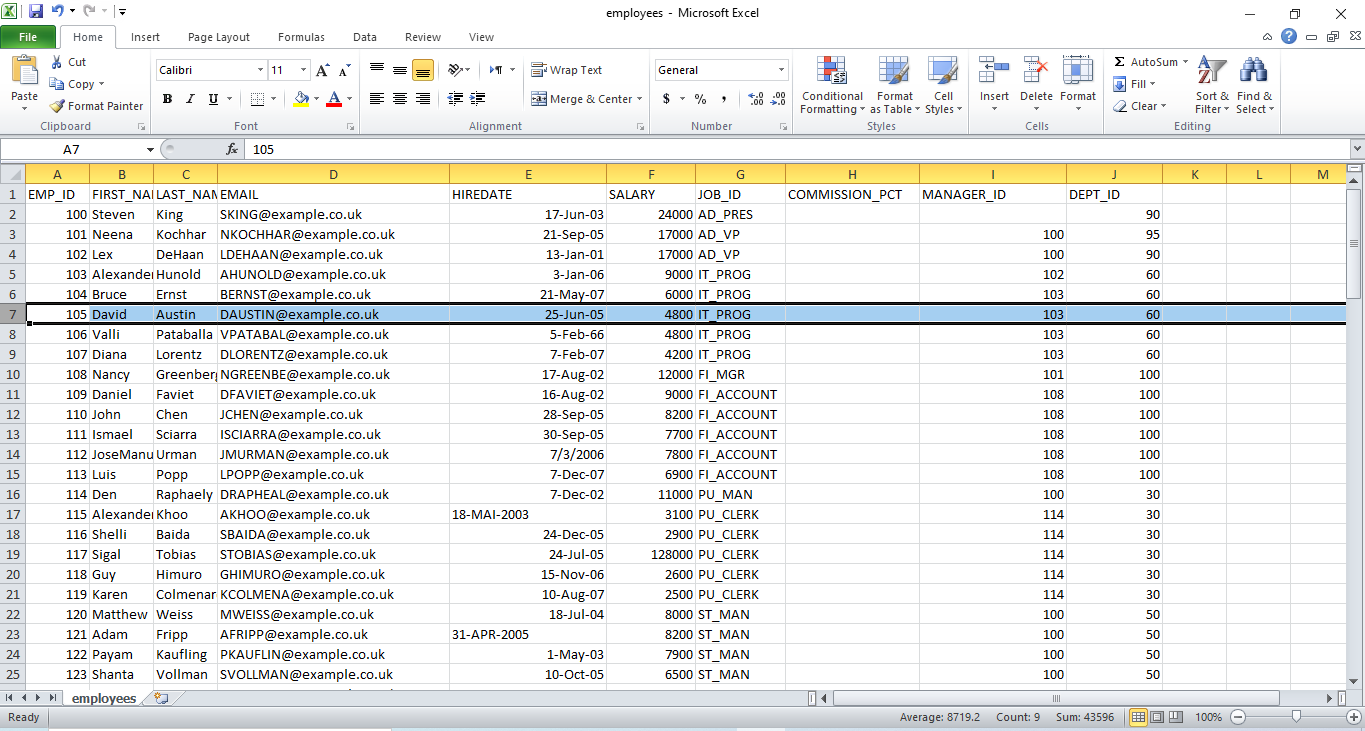
# 3. Evidence

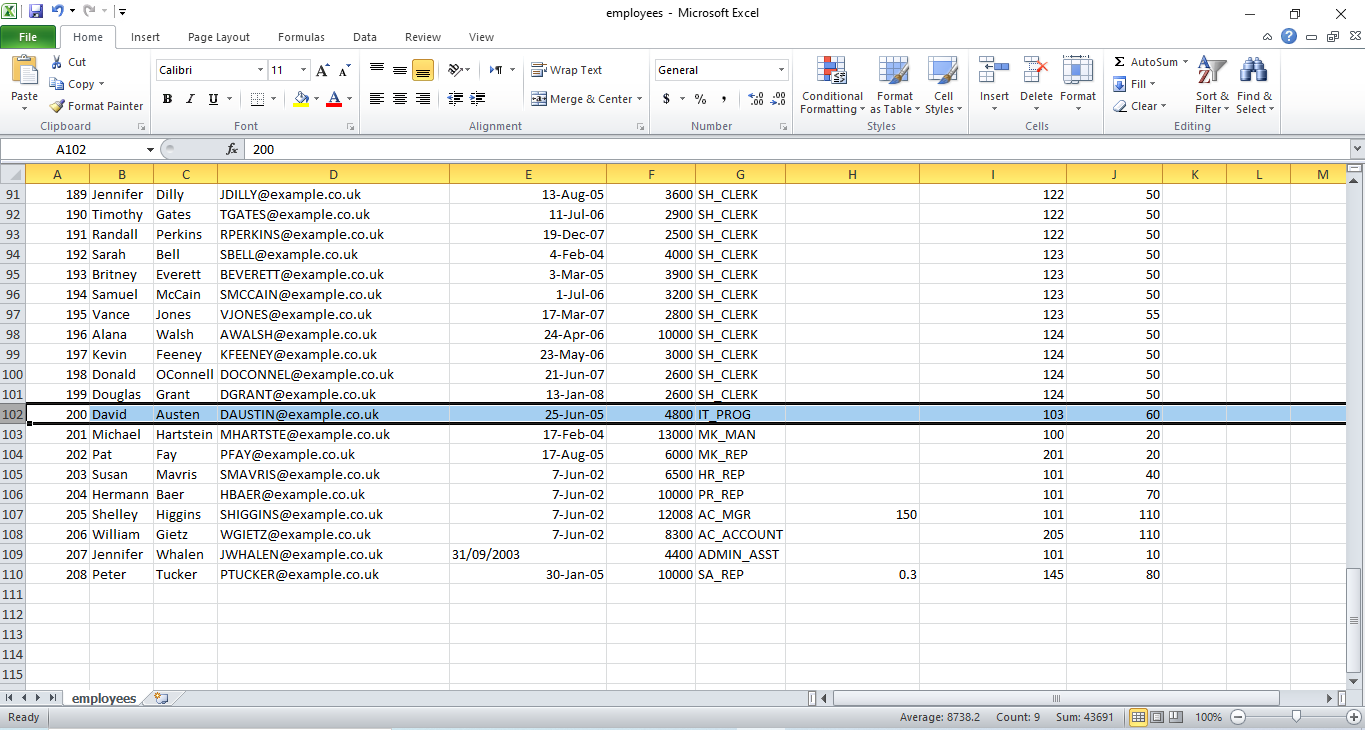
## Evidence to show the issue

First of all finding the missing data

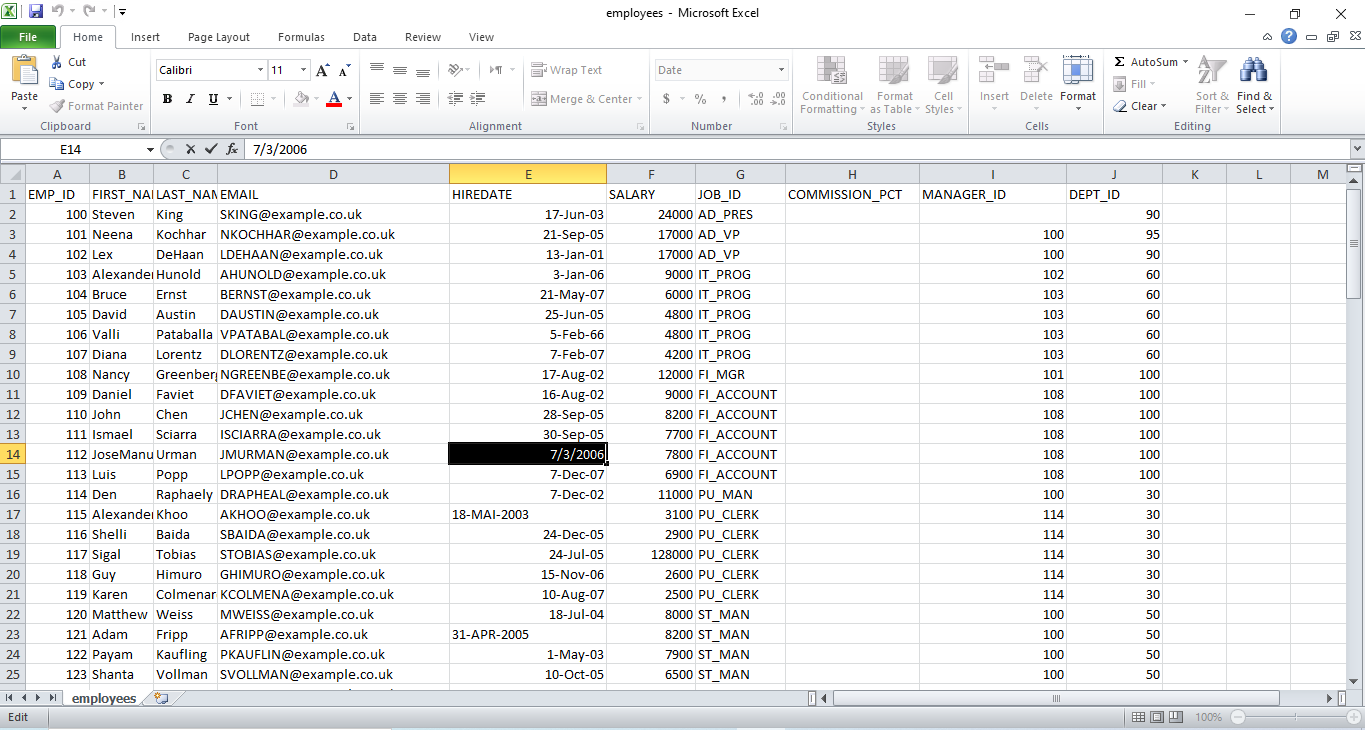


There is missing of data in Commission\_PCT

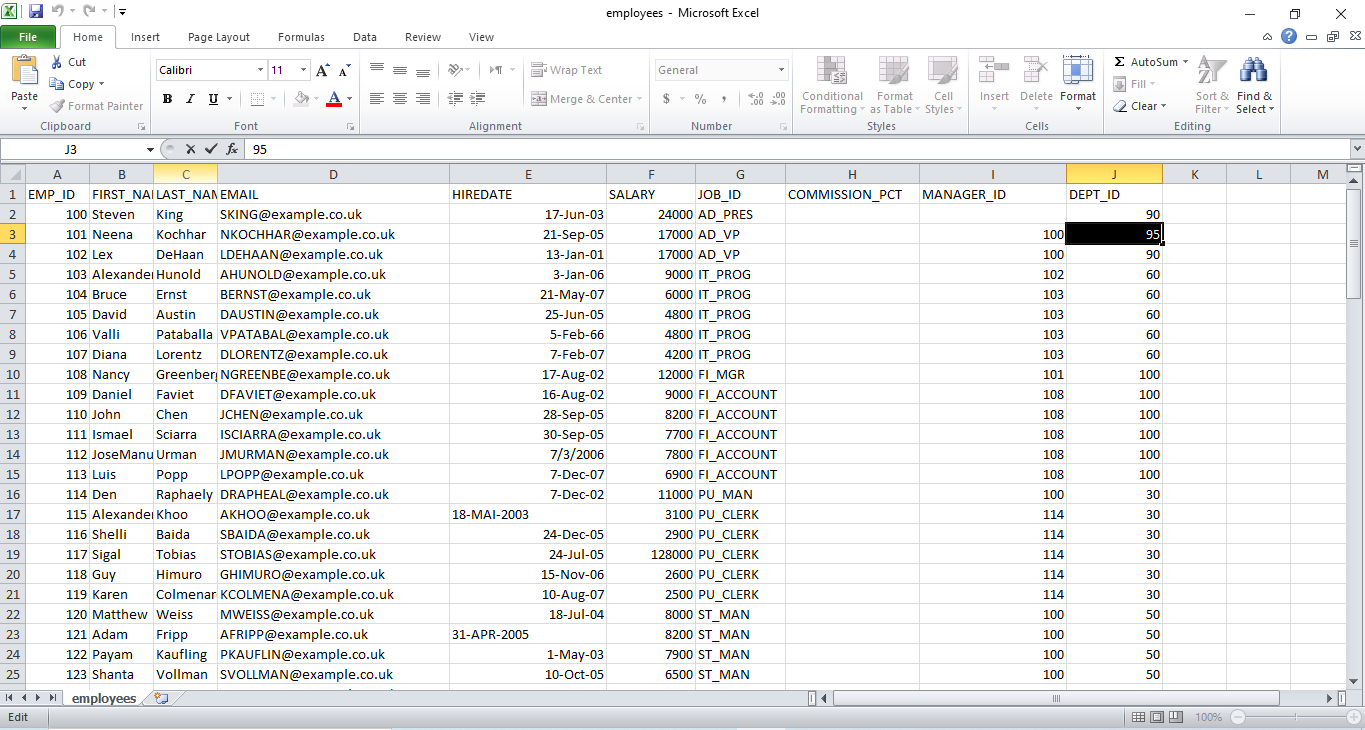
Duplicate Data



We can see that there is duplicate data in 102 and 7 rows.

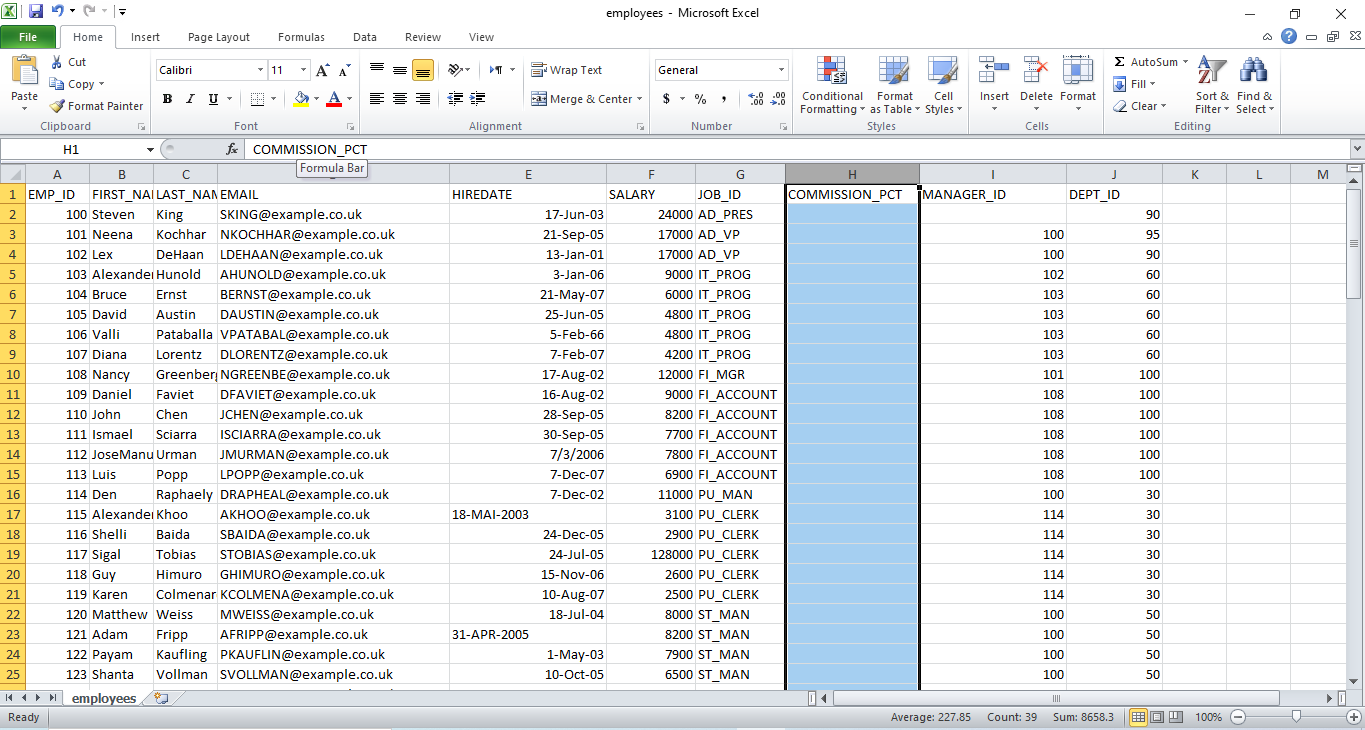
Inconsistent Data

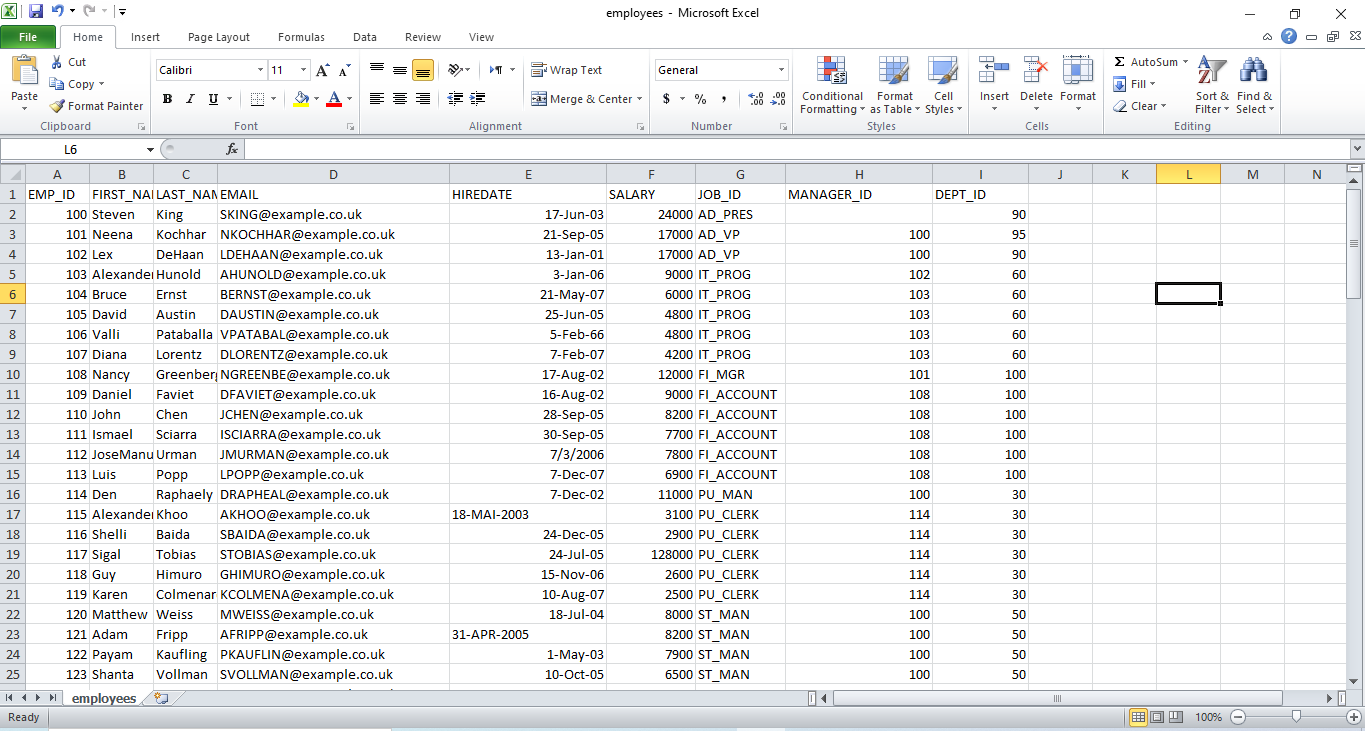
In row 14 we can see there is inconsistent data under Hiredate column. There is difference in date format with others date.

**Incorrect Data**

## Cleaning Data

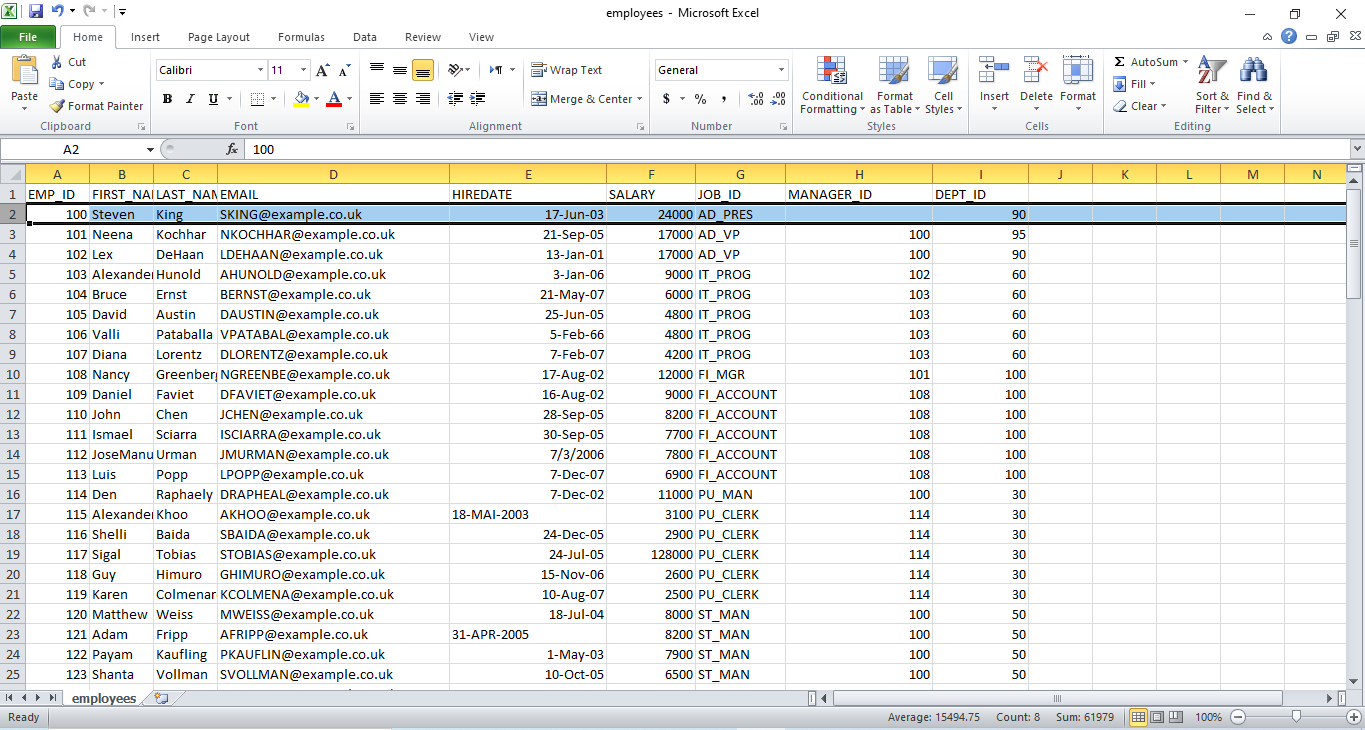
Removing the missing values Commision\_PCT

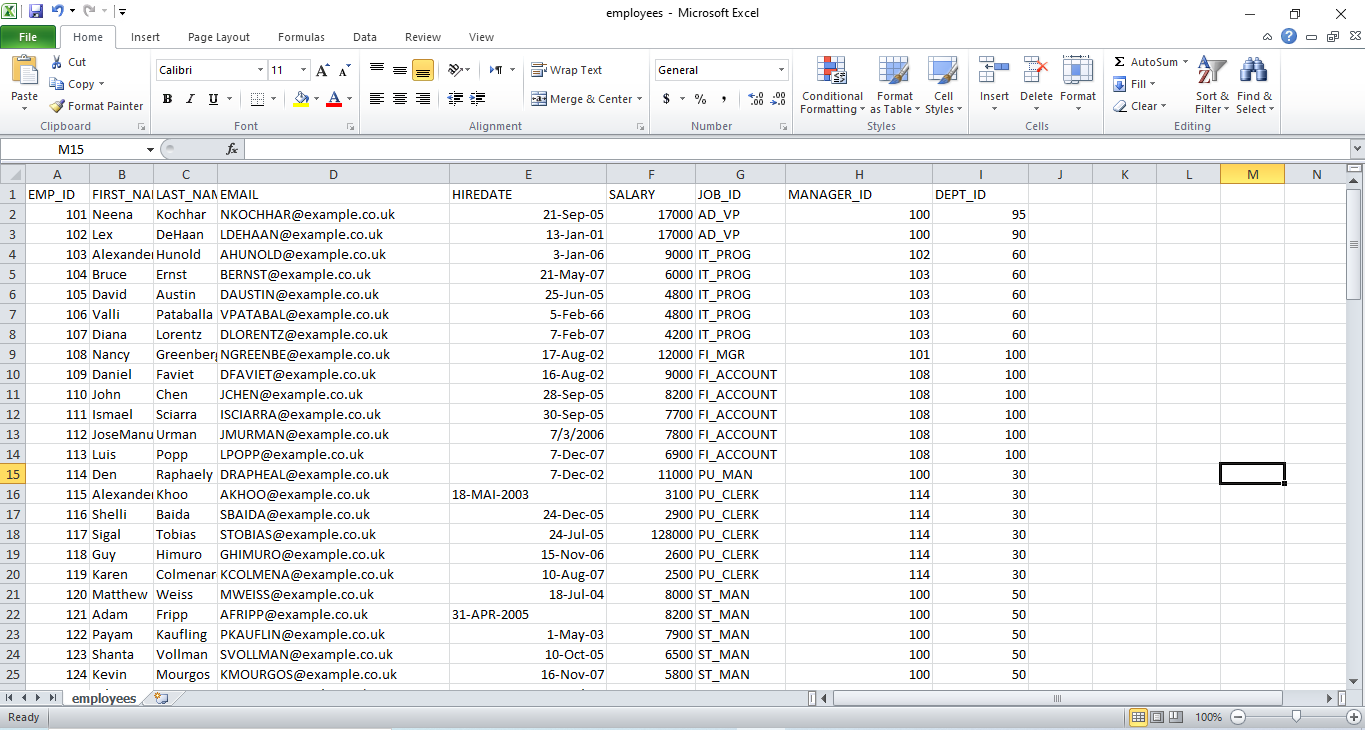
****Before:

****After:

We can see that there is no more Commision\_PCT column.

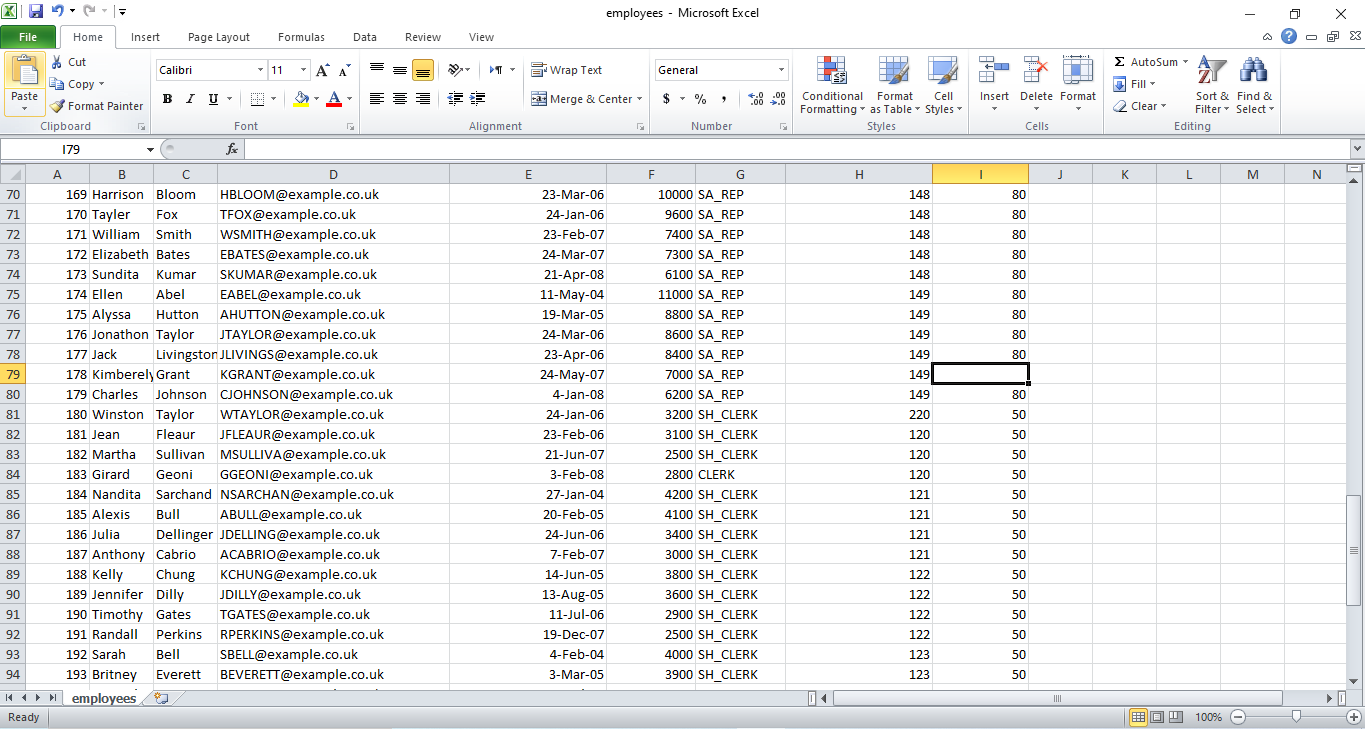
Delete the value cannot be replaced

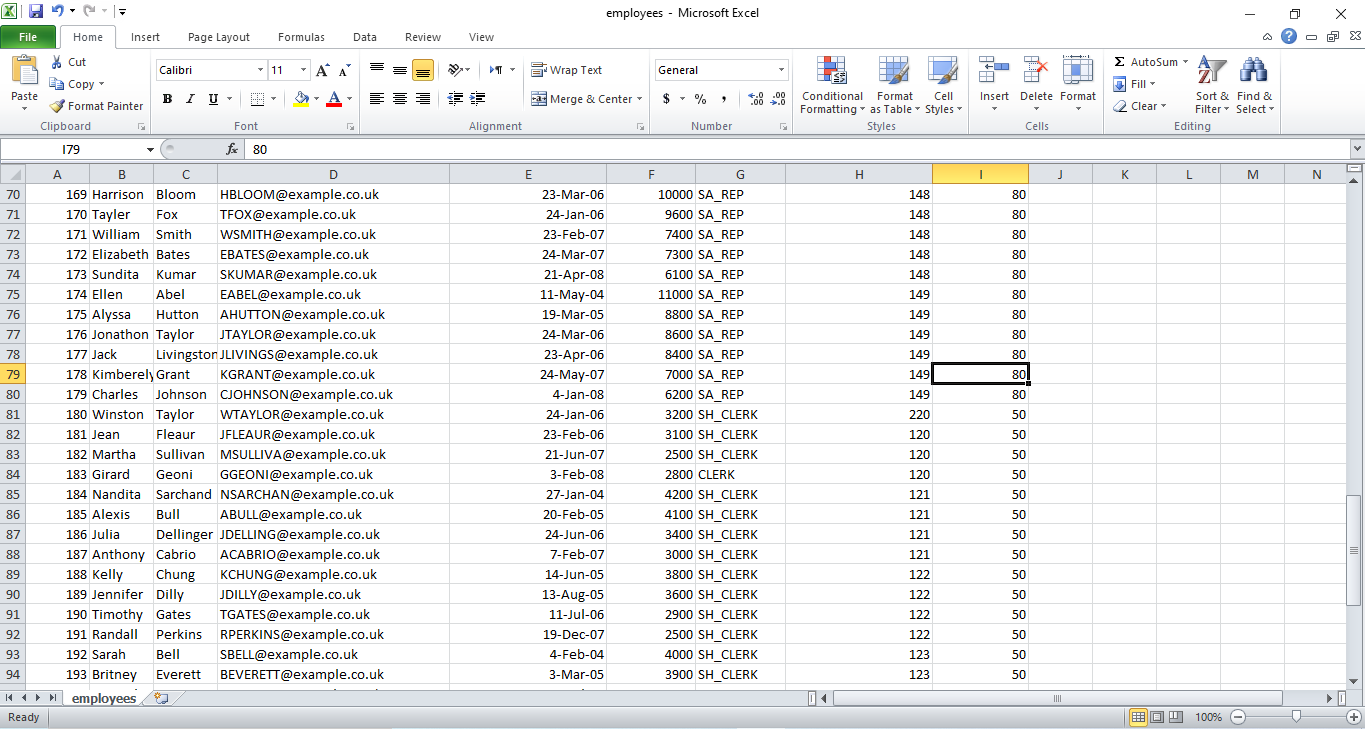
****Before:

****After

From row 2 First\_Name: steven is removed since it has no value in Manger\_ID column.

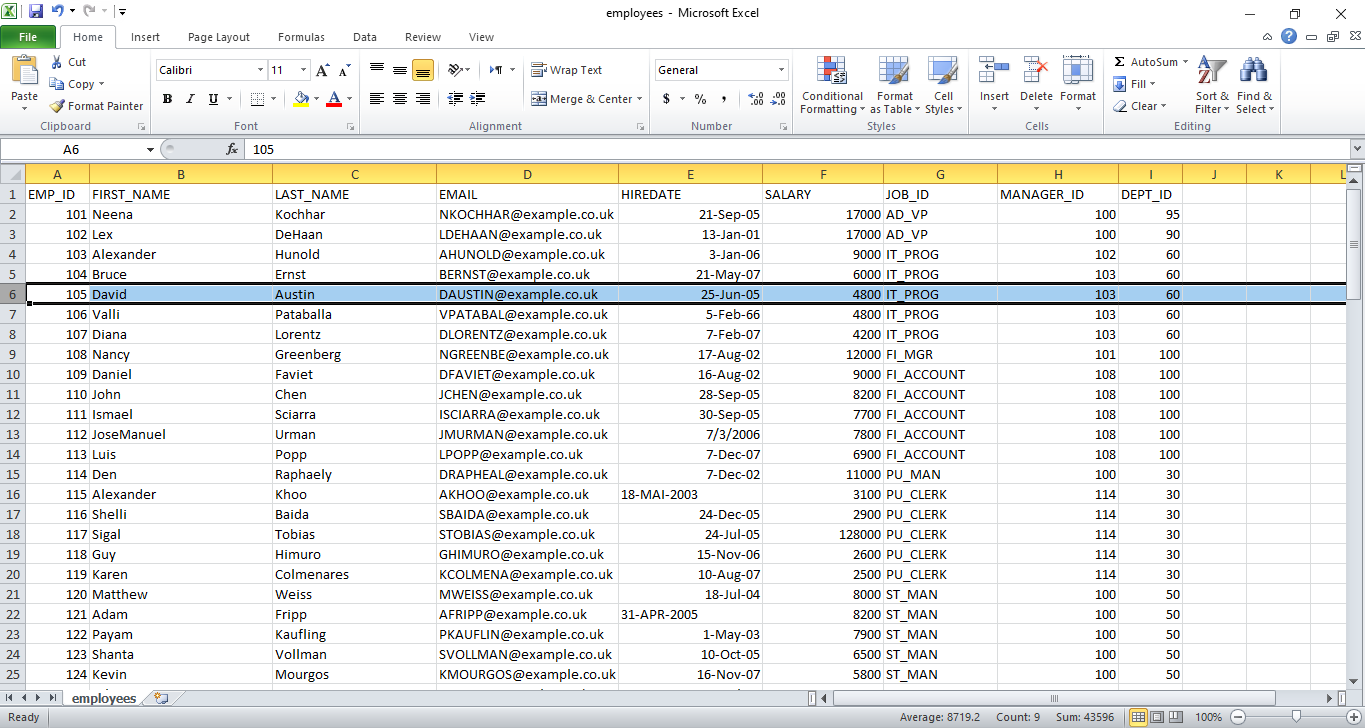
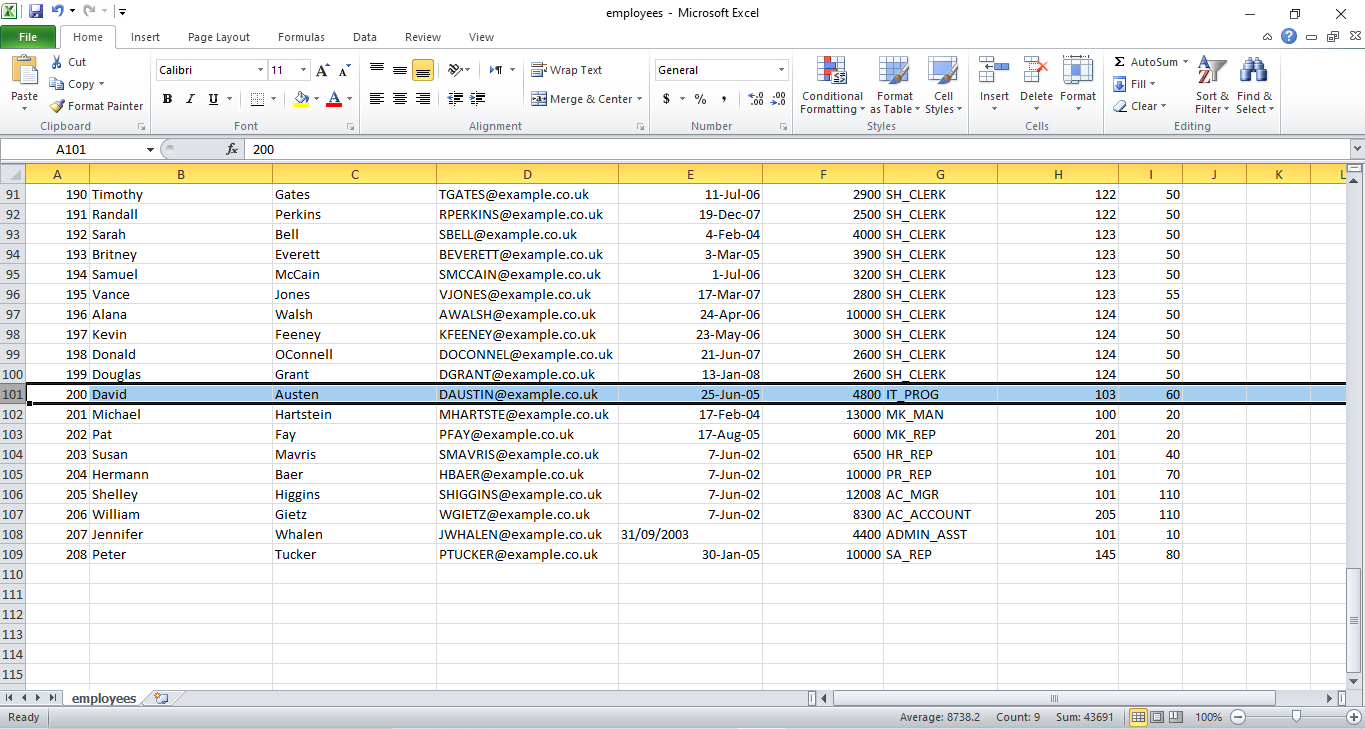
Putting Data in Empty Filed

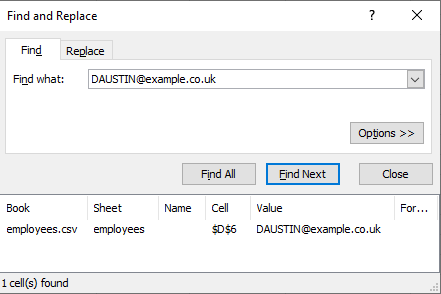
****Before:

****After:

As seen in the row 79, Dept\_ID column there is empty value, so in second figure, data is put in empty filed.

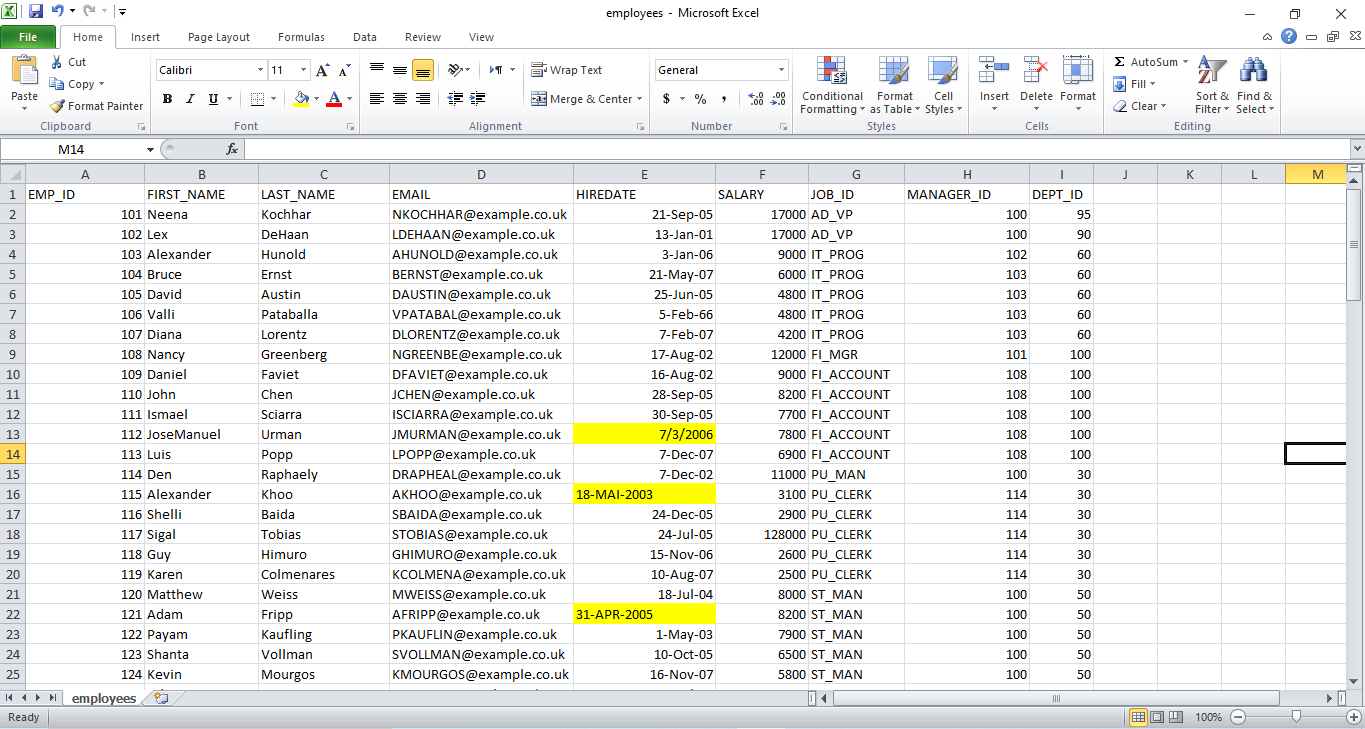
Duplicate Data

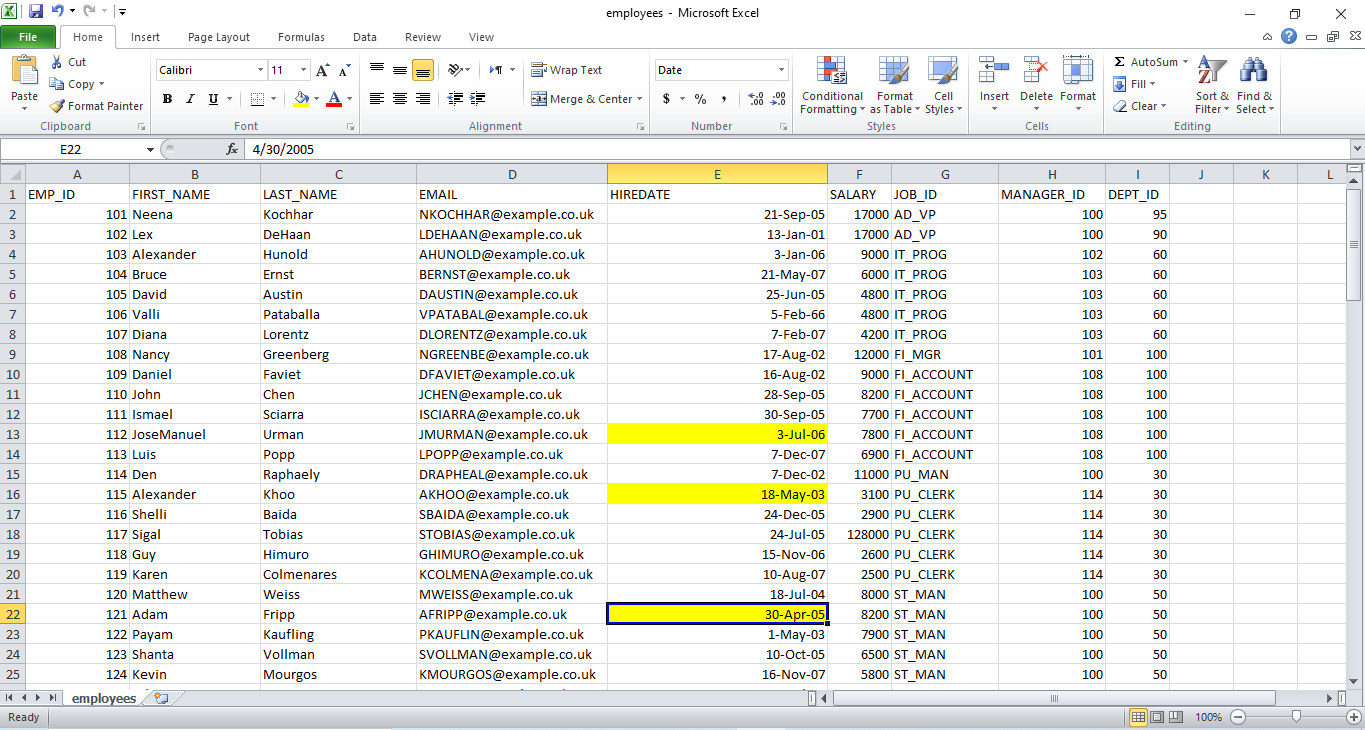
Before:

****After:

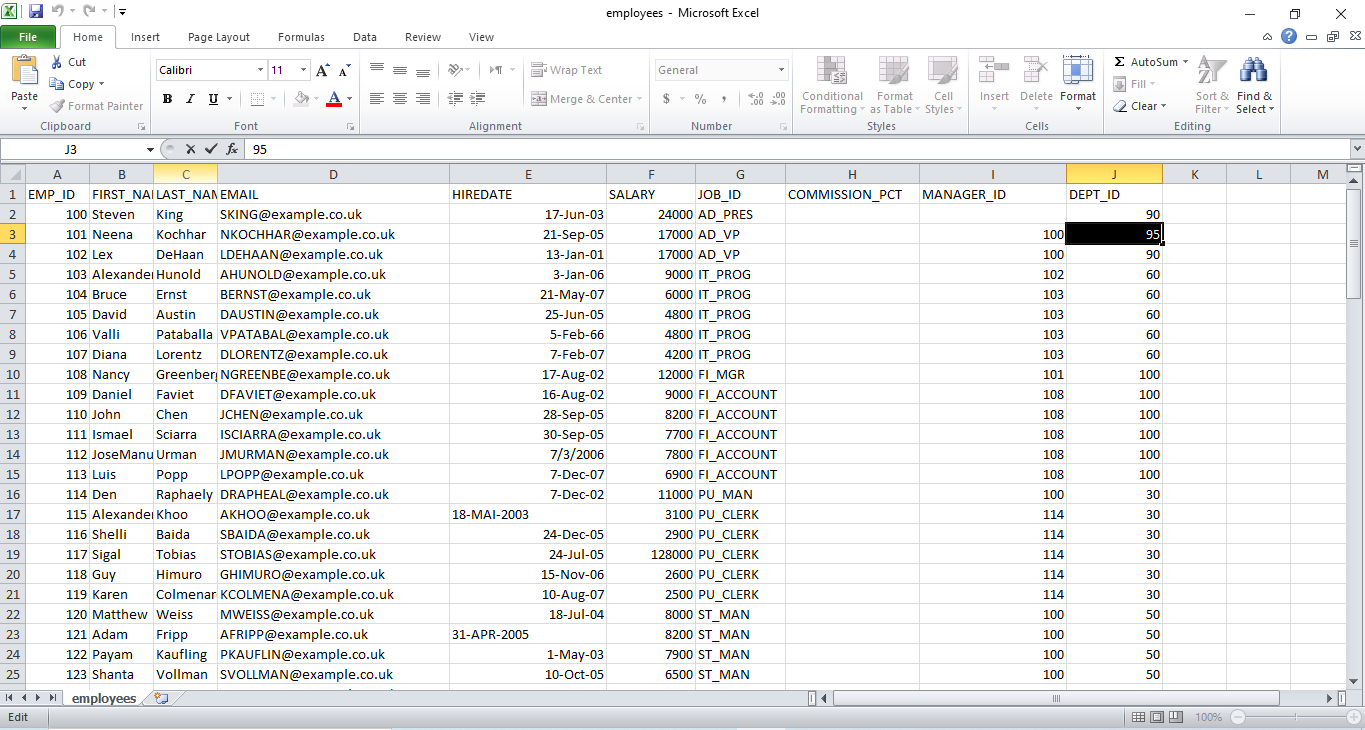
As we can see David Austin and David Austen has same email address therefore David Austen has been removed.

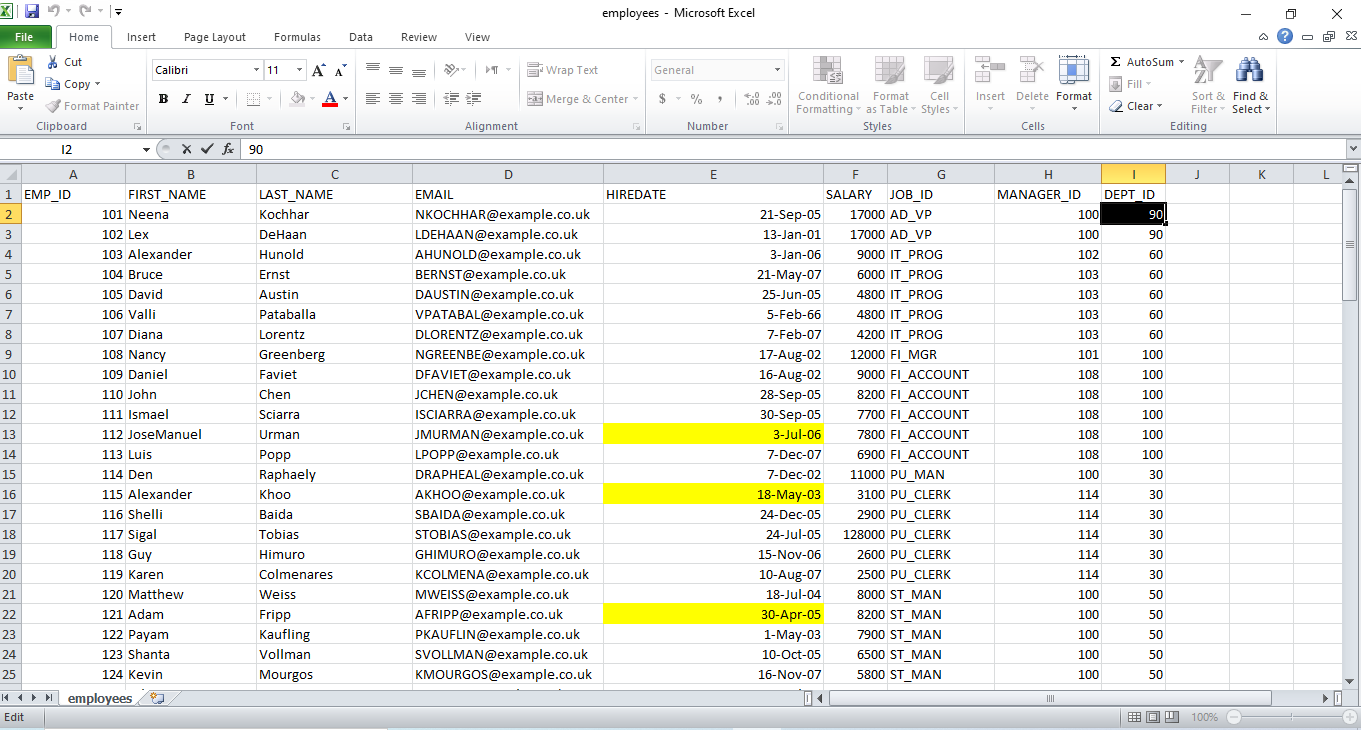
Inconsistent Data

****Before:

****After:

Incorrect Data:

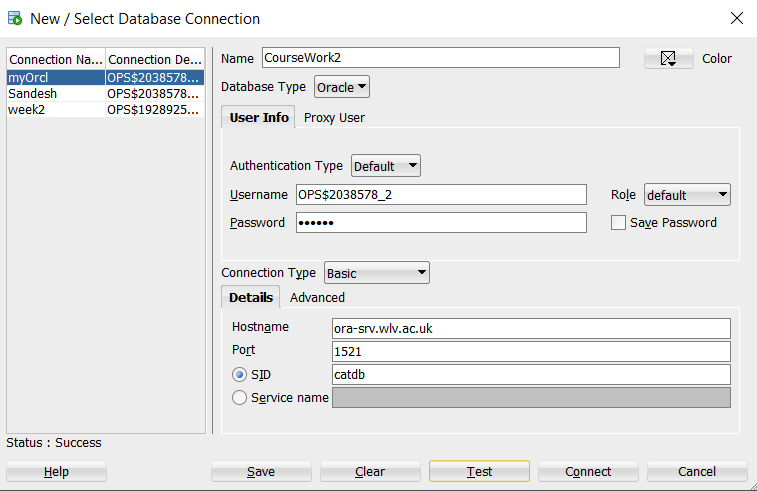
Before:

****After:

Neena Kochhar dept\_id should be 90 instead of 95 since her job id is ad\_vp as we can wee in the above image there is no depart id with 95.

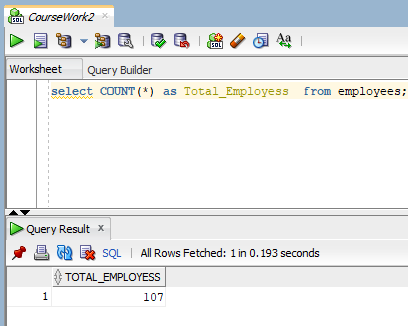
## Importing

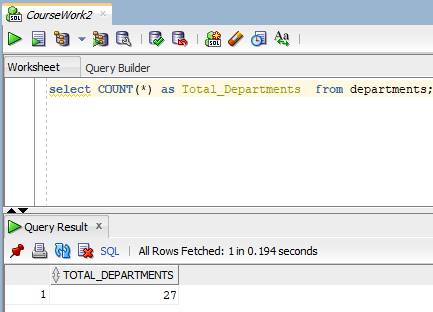
Creating new connection

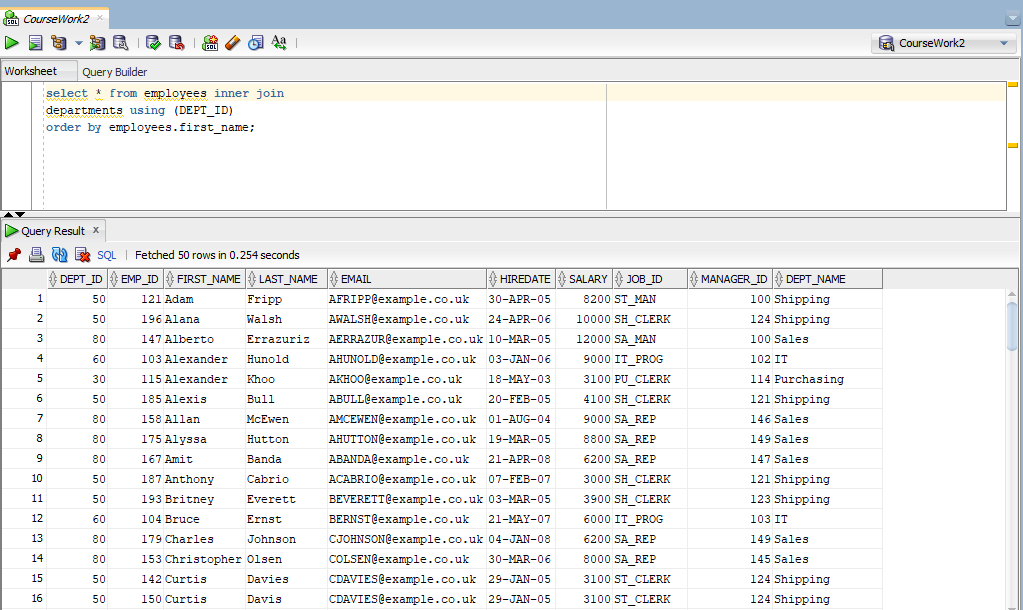


## SQL queries

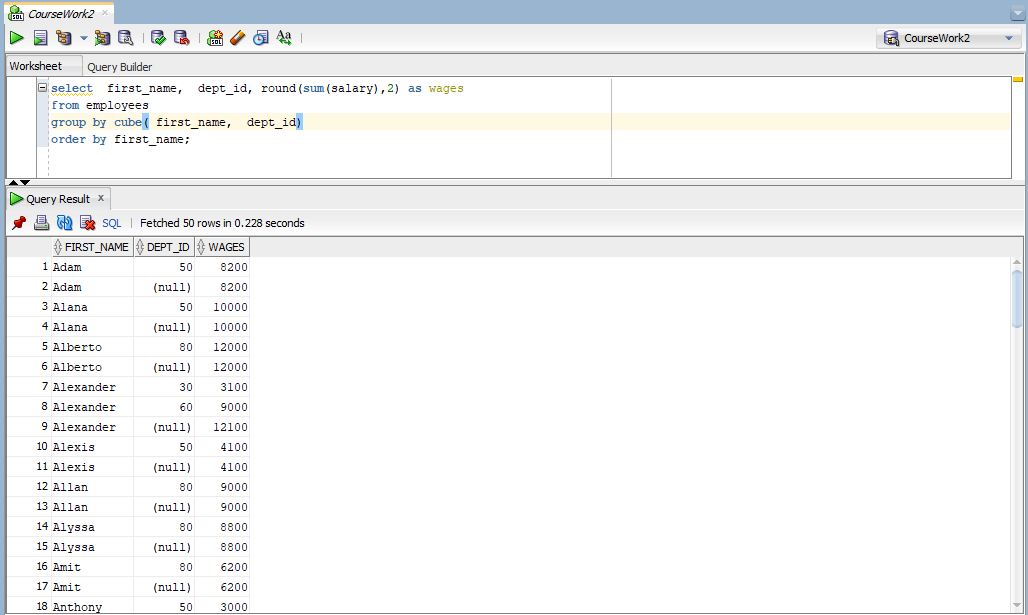
Counting total number of emp:



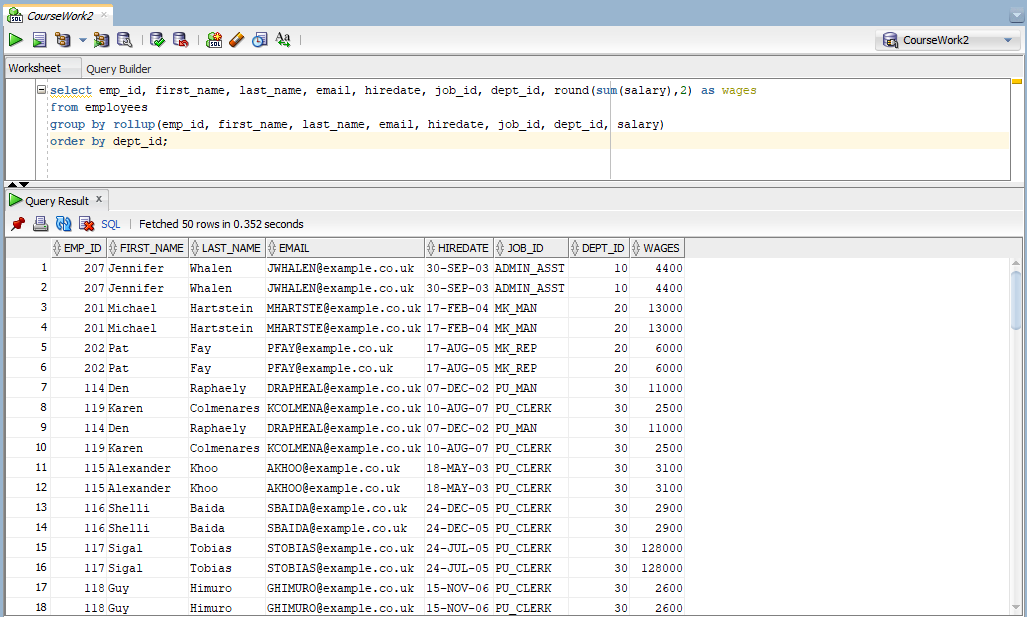
****Counting number of total dept

****Join Emp and Dep table:

Cube Command:

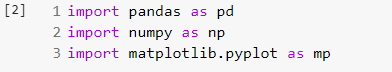
****

Rollup command



## Visualization

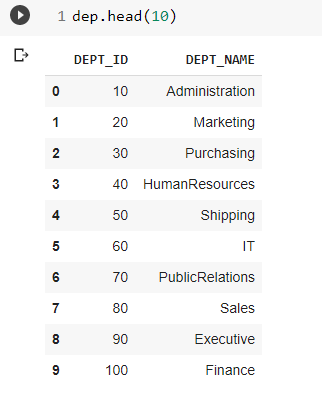
Importing required libraries

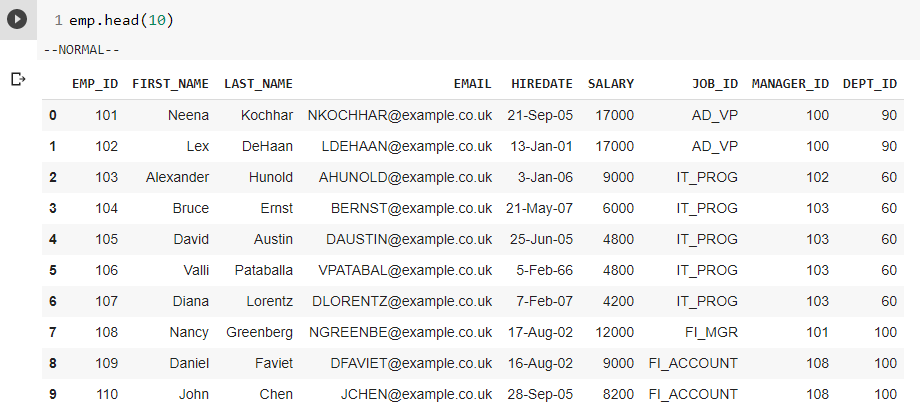


Importing files from drive and reading the files using pandas



Showing the first 10 data of dept

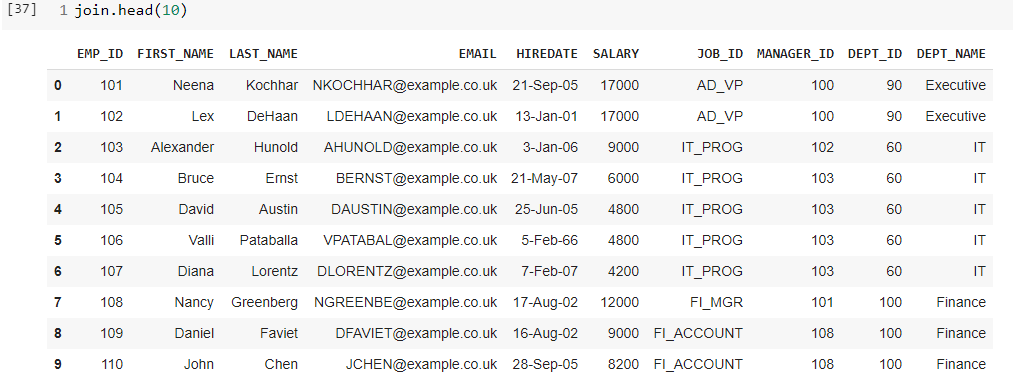


Showing the first 10 data of emp

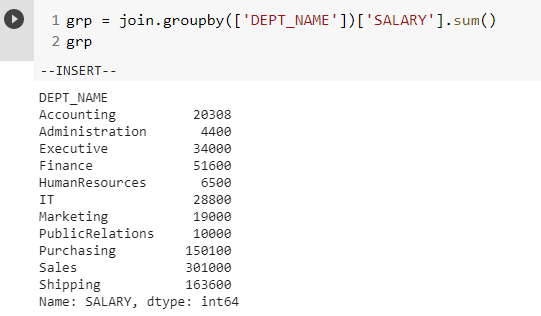
Based on dept\_id, emp and dept table are inner joined

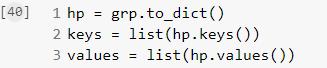


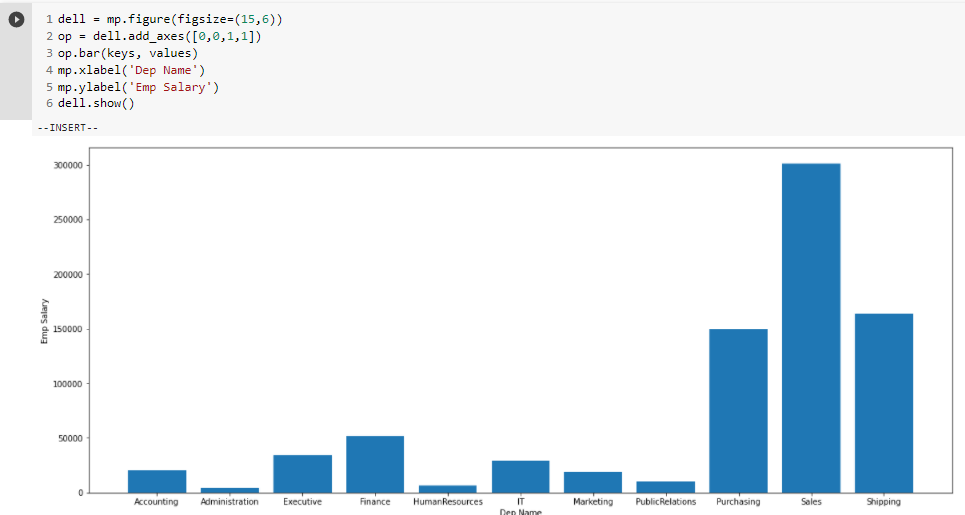
Displaying the first 10 data result after joining

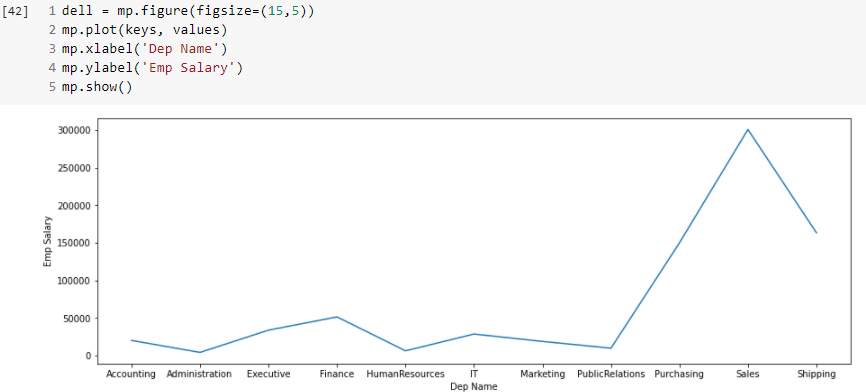


Summing the salary of emp and grouping by the dept\_name and displaying the result



Changing the grouping result into python dict

Visualizing the above data in bar diagram and line diagram respectively



# Bibliography

Kim, W. et al., 2003. A Taxonomy of Dirty Data. *Data mining and knowledge discovery,* 7(1), pp. 81-99.

Techopedia, 2020. *Techopedia.* [Online]   
Available at: https://www.techopedia.com/definition/14653/data-quality#:~:text=Techopedia%20Explains%20Data%20Quality,-Effective%20data%20quality&text=Completeness%3A%20Level%20at%20which%20desired,of%20various%20lists%20and%20mapping.  
[Accessed 22 4 2021].