UNIVERSITY PARTNER

UNIVERSITY OF WOLVERHAMPTON

HERALD COLLEGE KATHMANDU

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

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.

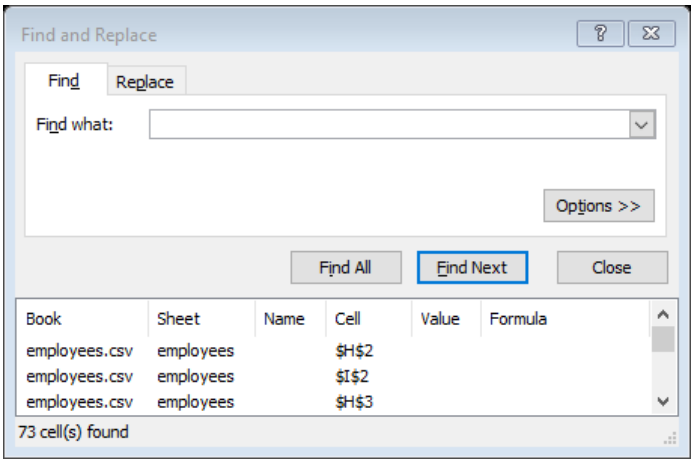
- Incorrect Data:

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

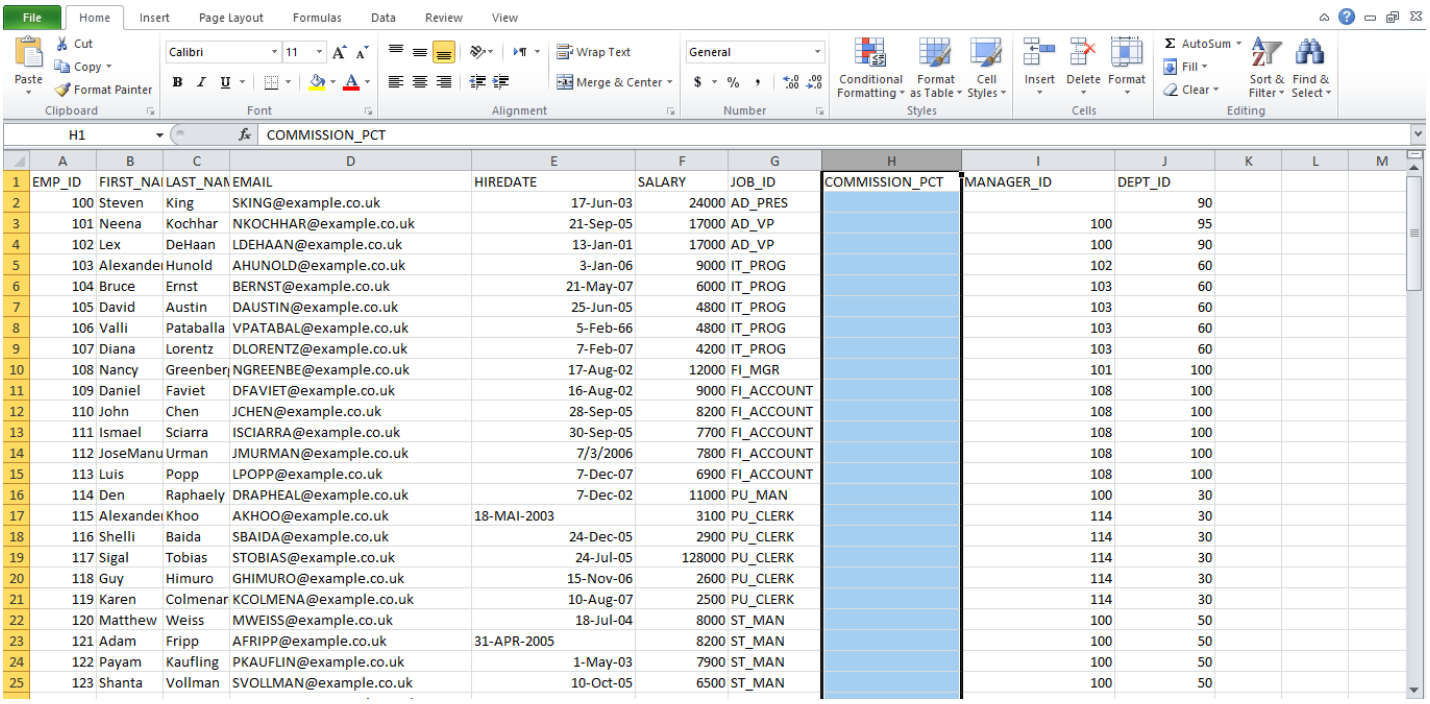
3. Evidence

Evidence to show the issue

First of all finding the missing data

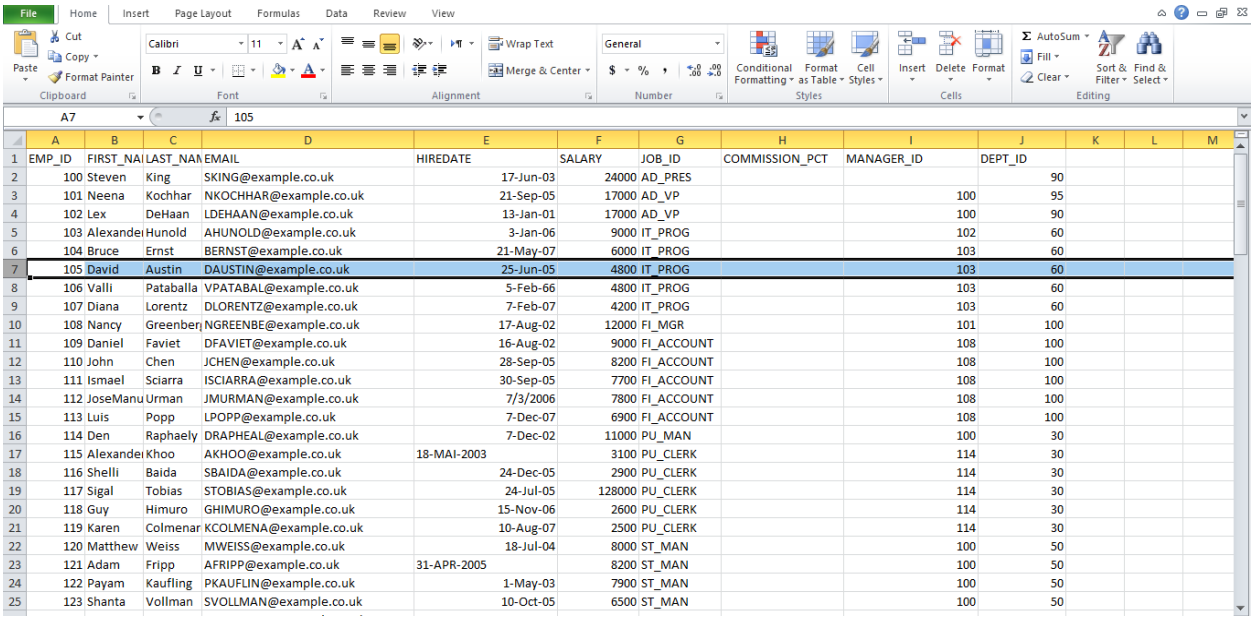


There is missing of data in Commission_PCT

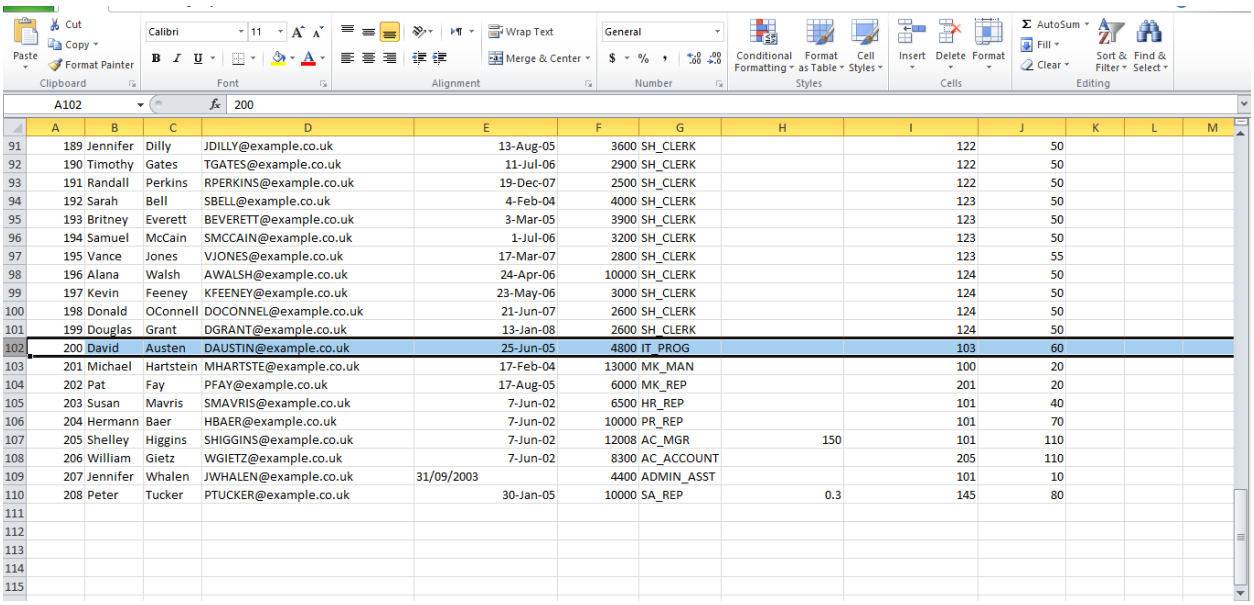


	A	B	C	D	E	F	G	H	I	J
1	EMP_ID	FIRST_NAI	LAST_NAI	EMAIL	HIREDATE	SALARY	JOB_ID	COMMISSION_PCT	MANAGER_ID	DEPT_ID
2	100	Steven	King	SKING@example.co.uk	17-Jun-03	24000	AD_PRES			90
3	101	Neena	Kochhar	NKOCHHAR@example.co.uk	21-Sep-05	17000	AD_VP		100	95
4	102	Lex	DeHaan	LDEHAAN@example.co.uk	13-Jan-01	17000	AD_VP		100	90
5	103	Alexandei	Hunold	AHUNOLD@example.co.uk	3-Jan-06	9000	IT_PROG		102	60
6	104	Bruce	Ernst	BERNST@example.co.uk	21-May-07	6000	IT_PROG		103	60
7	105	David	Austin	DAUSTIN@example.co.uk	25-Jun-05	4800	IT_PROG		103	60
8	106	Valli	Pataballa	VPATABAL@example.co.uk	5-Feb-66	4800	IT_PROG		103	60
9	107	Diana	Lorentz	DLORENTZ@example.co.uk	7-Feb-07	4200	IT_PROG		103	60
10	108	Nancy	Greenber	NGREENBE@example.co.uk	17-Aug-02	12000	FI_MGR		101	100
11	109	Daniel	Faviet	DFAVIET@example.co.uk	16-Aug-02	9000	FI_ACCOUNT		108	100
12	110	John	Chen	JCHEN@example.co.uk	28-Sep-05	8200	FI_ACCOUNT		108	100
13	111	Ismael	Sciarra	ISCIARRA@example.co.uk	30-Sep-05	7700	FI_ACCOUNT		108	100
14	112	JoseManu	Urman	JMURMAN@example.co.uk	7/3/2006	7800	FI_ACCOUNT		108	100
15	113	Luis	Popp	LPOPP@example.co.uk	7-Dec-07	6900	FI_ACCOUNT		108	100
16	114	Den	Raphaely	DRAPHEAL@example.co.uk	7-Dec-02	11000	PU_MAN		100	30
17	115	Alexandei	Khoo	AKHOO@example.co.uk	18-MAI-2003	3100	PU_CLERK		114	30
18	116	Shelli	Baida	SBAIDA@example.co.uk	24-Dec-05	2900	PU_CLERK		114	30
19	117	Sigal	Tobias	STOBIAS@example.co.uk	24-Jul-05	128000	PU_CLERK		114	30
20	118	Guy	Himuro	GHIMURO@example.co.uk	15-Nov-06	2600	PU_CLERK		114	30
21	119	Karen	Colmenar	KCOLMENA@example.co.uk	10-Aug-07	2500	PU_CLERK		114	30
22	120	Matthew	Weiss	MWEISS@example.co.uk	18-Jul-04	8000	ST_MAN		100	50
23	121	Adam	Fripp	AFRIPP@example.co.uk	31-APR-2005	8200	ST_MAN		100	50
24	122	Payam	Kaufling	PKAUFLIN@example.co.uk	1-May-03	7900	ST_MAN		100	50
25	123	Shanta	Vollman	SVOLLMAN@example.co.uk	10-Oct-05	6500	ST_MAN		100	50

Duplicate Data



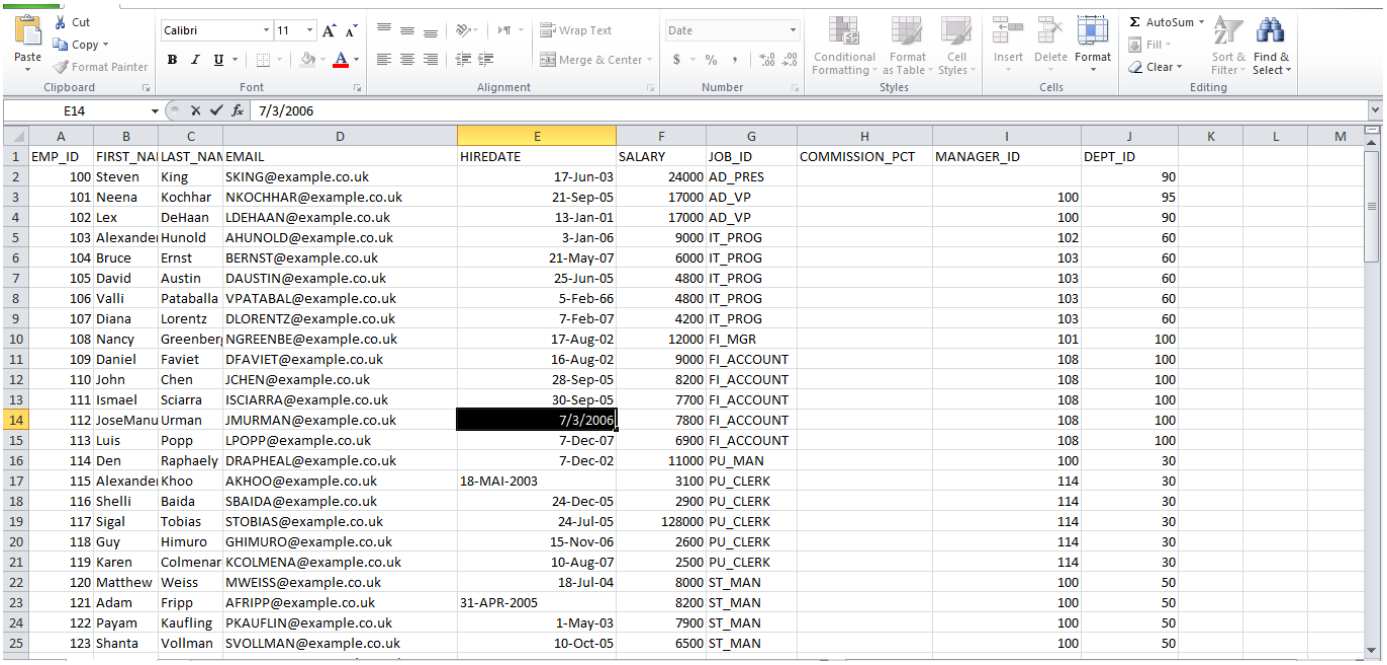
EMP_ID	FIRST_NAI	LAST_NAM	EMAIL	HIREDATE	SALARY	JOB_ID	COMMISSION_PCT	MANAGER_ID	DEPT_ID
100	Steven	King	SKING@example.co.uk	17-Jun-03	24000	AD_PRES			90
101	Neena	Kochhar	NKOCHHAR@example.co.uk	21-Sep-05	17000	AD_VP		100	95
102	Lex	DeHaan	LDEHAAN@example.co.uk	13-Jan-01	17000	AD_VP		100	90
103	Alexander	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	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
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	JoseManu	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	Colmenar	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
122	Payam	Kaufling	PKAUFLIN@example.co.uk	1-May-03	7900	ST_MAN		100	50
123	Shanta	Vollman	SVOLLMAN@example.co.uk	10-Oct-05	6500	ST_MAN		100	50



189	Jennifer	Dilly	JDILLY@example.co.uk	13-Aug-05	3600	SH_CLERK		122	50
190	Timothy	Gates	TGATES@example.co.uk	11-Jul-06	2900	SH_CLERK		122	50
191	Randall	Perkins	RPERKINS@example.co.uk	19-Dec-07	2500	SH_CLERK		122	50
192	Sarah	Bell	SBELL@example.co.uk	4-Feb-04	4000	SH_CLERK		123	50
193	Britney	Everett	BEVERETT@example.co.uk	3-Mar-05	3900	SH_CLERK		123	50
194	Samuel	McCain	SMCCAIN@example.co.uk	1-Jul-06	3200	SH_CLERK		123	50
195	Vance	Jones	VJONES@example.co.uk	17-Mar-07	2800	SH_CLERK		123	55
196	Alana	Walsh	AWALSH@example.co.uk	24-Apr-06	10000	SH_CLERK		124	50
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
204	Hermann	Baer	HBAER@example.co.uk	7-Jun-02	10000	PR_REP		101	70
205	Shelley	Higgins	SHIGGINS@example.co.uk	7-Jun-02	12008	AC_MGR	150	101	110
206	William	Gietz	WGIETZ@example.co.uk	7-Jun-02	8300	AC_ACCOUNT		205	110
207	Jennifer	Whalen	JWHALEN@example.co.uk	31/09/2003	4400	ADMIN_ASST		101	10
208	Peter	Tucker	PTUCKER@example.co.uk	30-Jan-05	10000	SA_REP	0.3	145	80

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

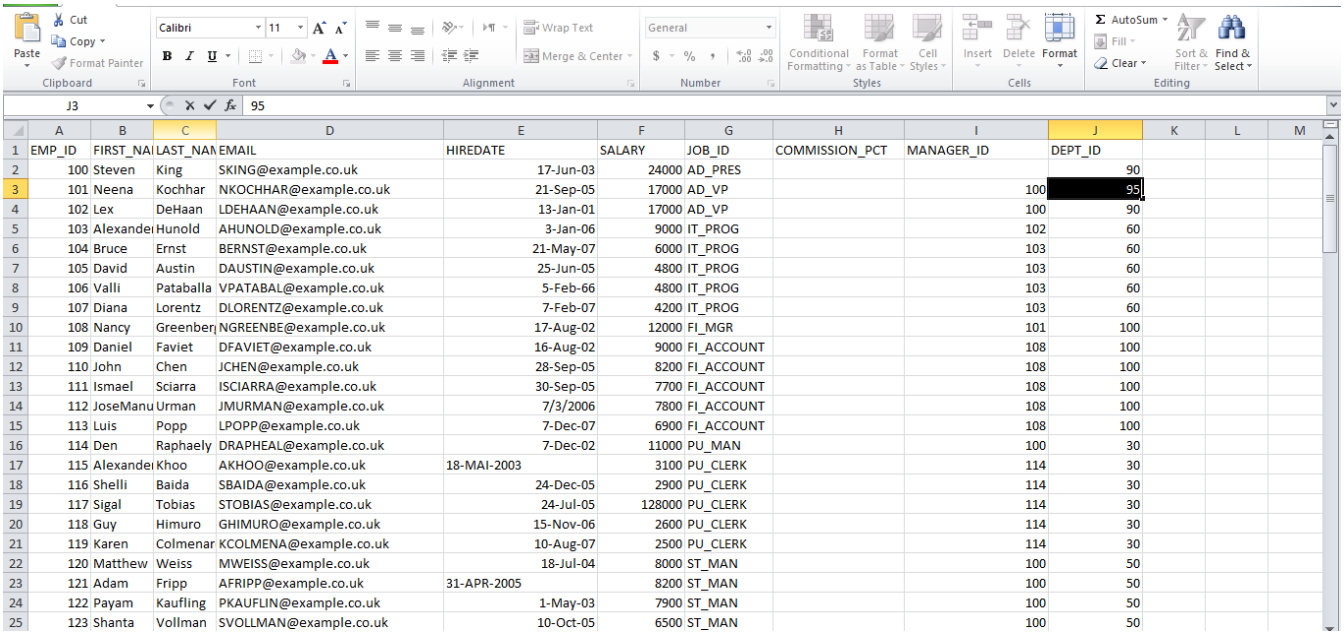
Inconsistent Data



	A	B	C	D	E	F	G	H	I	J
1	EMP_ID	FIRST_NAI	LAST_NAM	EMAIL	HIREDATE	SALARY	JOB_ID	COMMISSION_PCT	MANAGER_ID	DEPT_ID
2	100	Steven	King	SKING@example.co.uk	17-Jun-03	24000	AD_PRES			90
3	101	Neena	Kochhar	NKOCHHAR@example.co.uk	21-Sep-05	17000	AD_VP		100	95
4	102	Lex	DeHaan	LDEHAAN@example.co.uk	13-Jan-01	17000	AD_VP		100	90
5	103	Alexander	Hunold	AHUNOLD@example.co.uk	3-Jan-06	9000	IT_PROG		102	60
6	104	Bruce	Ernst	BERNST@example.co.uk	21-May-07	6000	IT_PROG		103	60
7	105	David	Austin	DAUSTIN@example.co.uk	25-Jun-05	4800	IT_PROG		103	60
8	106	Valli	Pataballa	VPATABAL@example.co.uk	5-Feb-66	4800	IT_PROG		103	60
9	107	Diana	Lorentz	DLORENTZ@example.co.uk	7-Feb-07	4200	IT_PROG		103	60
10	108	Nancy	Greenber	NGREENBE@example.co.uk	17-Aug-02	12000	FI_MGR		101	100
11	109	Daniel	Faviet	DFAVIET@example.co.uk	16-Aug-02	9000	FI_ACCOUNT		108	100
12	110	John	Chen	JCHEN@example.co.uk	28-Sep-05	8200	FI_ACCOUNT		108	100
13	111	Ismael	Sciarra	ISCIARRA@example.co.uk	30-Sep-05	7700	FI_ACCOUNT		108	100
14	112	JoseManu	Urman	JMURMAN@example.co.uk	7/3/2006	7800	FI_ACCOUNT		108	100
15	113	Luis	Popp	LPOPP@example.co.uk	7-Dec-07	6900	FI_ACCOUNT		108	100
16	114	Den	Raphaely	DRAPHEAL@example.co.uk	7-Dec-02	11000	PU_MAN		100	30
17	115	Alexander	Khoo	AKHOO@example.co.uk	18-MAI-2003	3100	PU_CLERK		114	30
18	116	Shelli	Baida	SBAIDA@example.co.uk	24-Dec-05	2900	PU_CLERK		114	30
19	117	Sigal	Tobias	STOBIAS@example.co.uk	24-Jul-05	128000	PU_CLERK		114	30
20	118	Guy	Himuro	GHIMURO@example.co.uk	15-Nov-06	2600	PU_CLERK		114	30
21	119	Karen	Colmenar	KCOLMENA@example.co.uk	10-Aug-07	2500	PU_CLERK		114	30
22	120	Matthew	Weiss	MWEISS@example.co.uk	18-Jul-04	8000	ST_MAN		100	50
23	121	Adam	Fripp	AFRIPP@example.co.uk	31-APR-2005	8200	ST_MAN		100	50
24	122	Payam	Kaufling	PKAUFLIN@example.co.uk	1-May-03	7900	ST_MAN		100	50
25	123	Shanta	Vollman	SVOLLMAN@example.co.uk	10-Oct-05	6500	ST_MAN		100	50

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

Incorrect Data

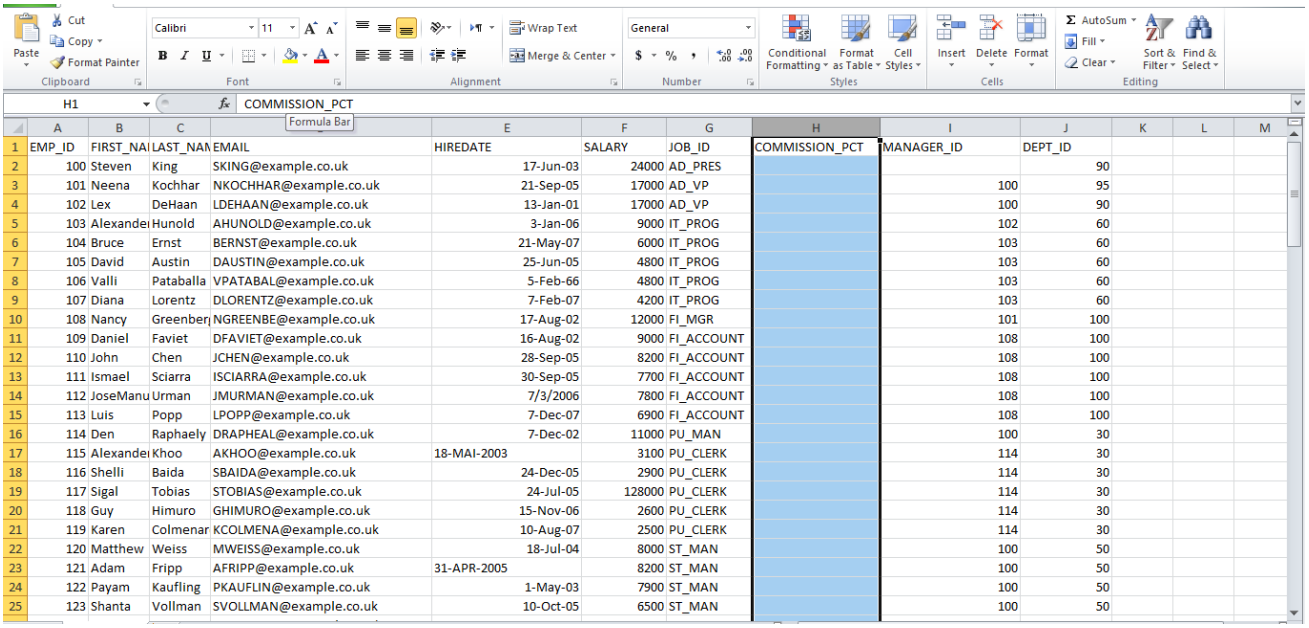


	A	B	C	D	E	F	G	H	I	J
1	EMP_ID	FIRST_NAI	LAST_NAM	EMAIL	HIREDATE	SALARY	JOB_ID	COMMISSION_PCT	MANAGER_ID	DEPT_ID
2	100	Steven	King	SKING@example.co.uk	17-Jun-03	24000	AD_PRES			90
3	101	Neena	Kochhar	NKOCHHAR@example.co.uk	21-Sep-05	17000	AD_VP		100	95
4	102	Lex	DeHaan	LDEHAAN@example.co.uk	13-Jan-01	17000	AD_VP		100	90
5	103	Alexander	Hunold	AHUNOLD@example.co.uk	3-Jan-06	9000	IT_PROG		102	60
6	104	Bruce	Ernst	BERNST@example.co.uk	21-May-07	6000	IT_PROG		103	60
7	105	David	Austin	DAUSTIN@example.co.uk	25-Jun-05	4800	IT_PROG		103	60
8	106	Valli	Pataballa	VPATABAL@example.co.uk	5-Feb-66	4800	IT_PROG		103	60
9	107	Diana	Lorentz	DLORENTZ@example.co.uk	7-Feb-07	4200	IT_PROG		103	60
10	108	Nancy	Greenber	NGREENBE@example.co.uk	17-Aug-02	12000	FI_MGR		101	100
11	109	Daniel	Faviet	DFAVIET@example.co.uk	16-Aug-02	9000	FI_ACCOUNT		108	100
12	110	John	Chen	JCHEN@example.co.uk	28-Sep-05	8200	FI_ACCOUNT		108	100
13	111	Ismael	Sciarra	ISCIARRA@example.co.uk	30-Sep-05	7700	FI_ACCOUNT		108	100
14	112	JoseManu	Urman	JMURMAN@example.co.uk	7/3/2006	7800	FI_ACCOUNT		108	100
15	113	Luis	Popp	LPOPP@example.co.uk	7-Dec-07	6900	FI_ACCOUNT		108	100
16	114	Den	Raphaely	DRAPHEAL@example.co.uk	7-Dec-02	11000	PU_MAN		100	30
17	115	Alexander	Khoo	AKHOO@example.co.uk	18-MAI-2003	3100	PU_CLERK		114	30
18	116	Shelli	Baida	SBAIDA@example.co.uk	24-Dec-05	2900	PU_CLERK		114	30
19	117	Sigal	Tobias	STOBIAS@example.co.uk	24-Jul-05	128000	PU_CLERK		114	30
20	118	Guy	Himuro	GHIMURO@example.co.uk	15-Nov-06	2600	PU_CLERK		114	30
21	119	Karen	Colmenar	KCOLMENA@example.co.uk	10-Aug-07	2500	PU_CLERK		114	30
22	120	Matthew	Weiss	MWEISS@example.co.uk	18-Jul-04	8000	ST_MAN		100	50
23	121	Adam	Fripp	AFRIPP@example.co.uk	31-APR-2005	8200	ST_MAN		100	50
24	122	Payam	Kaufling	PKAUFLIN@example.co.uk	1-May-03	7900	ST_MAN		100	50
25	123	Shanta	Vollman	SVOLLMAN@example.co.uk	10-Oct-05	6500	ST_MAN		100	50

Cleaning Data

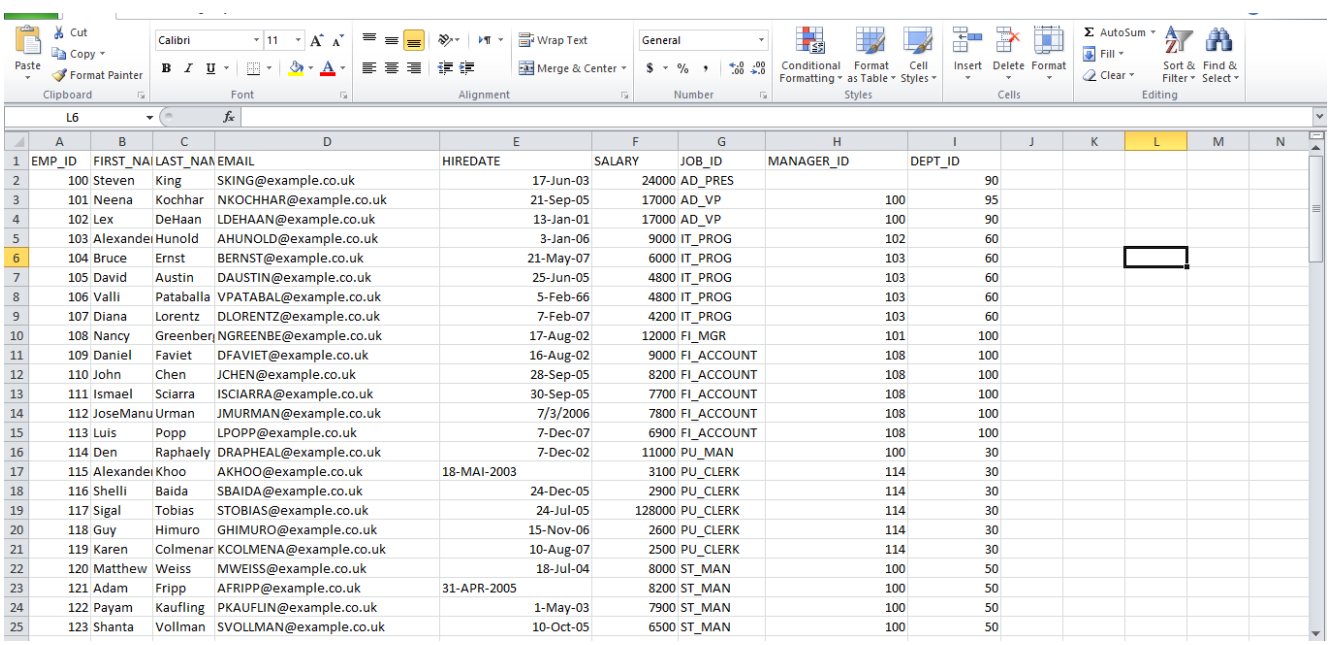
Removing the missing values Commision_PCT

Before:



EMP_ID	FIRST_NAI	LAST_NAM	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	Alexandei	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	JoseManu	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	Alexandei	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	Colmenar	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
122	Payam	Kaufling	PKAUFLIN@example.co.uk	1-May-03	7900	ST_MAN		100	50
123	Shanta	Vollman	SVOLLMAN@example.co.uk	10-Oct-05	6500	ST_MAN		100	50

After:

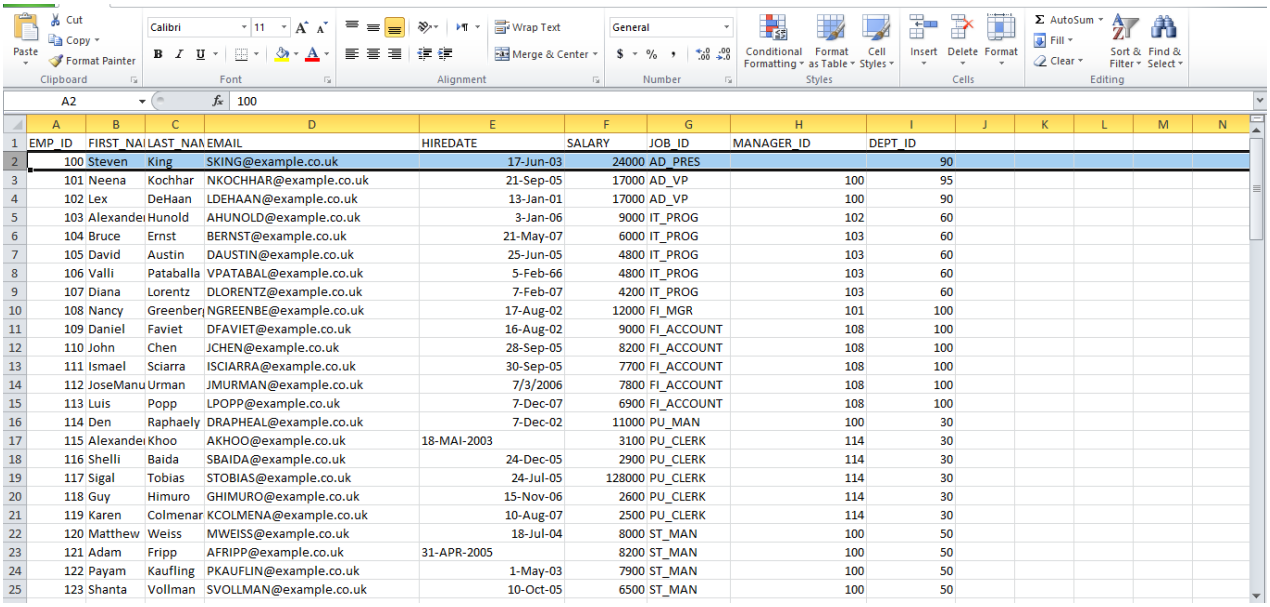


EMP_ID	FIRST_NAI	LAST_NAM	EMAIL	HIREDATE	SALARY	JOB_ID	MANAGER_ID	DEPT_ID
100	Steven	King	SKING@example.co.uk	17-Jun-03	24000	AD_PRES		90
101	Neena	Kochhar	NKOCHHAR@example.co.uk	21-Sep-05	17000	AD_VP	100	95
102	Lex	DeHaan	LDEHAAN@example.co.uk	13-Jan-01	17000	AD_VP	100	90
103	Alexandei	Hunold	AHUNOLD@example.co.uk	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	JoseManu	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	Alexandei	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	Colmenar	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
122	Payam	Kaufling	PKAUFLIN@example.co.uk	1-May-03	7900	ST_MAN	100	50
123	Shanta	Vollman	SVOLLMAN@example.co.uk	10-Oct-05	6500	ST_MAN	100	50

We can see that there is no more Commision_PCT column.

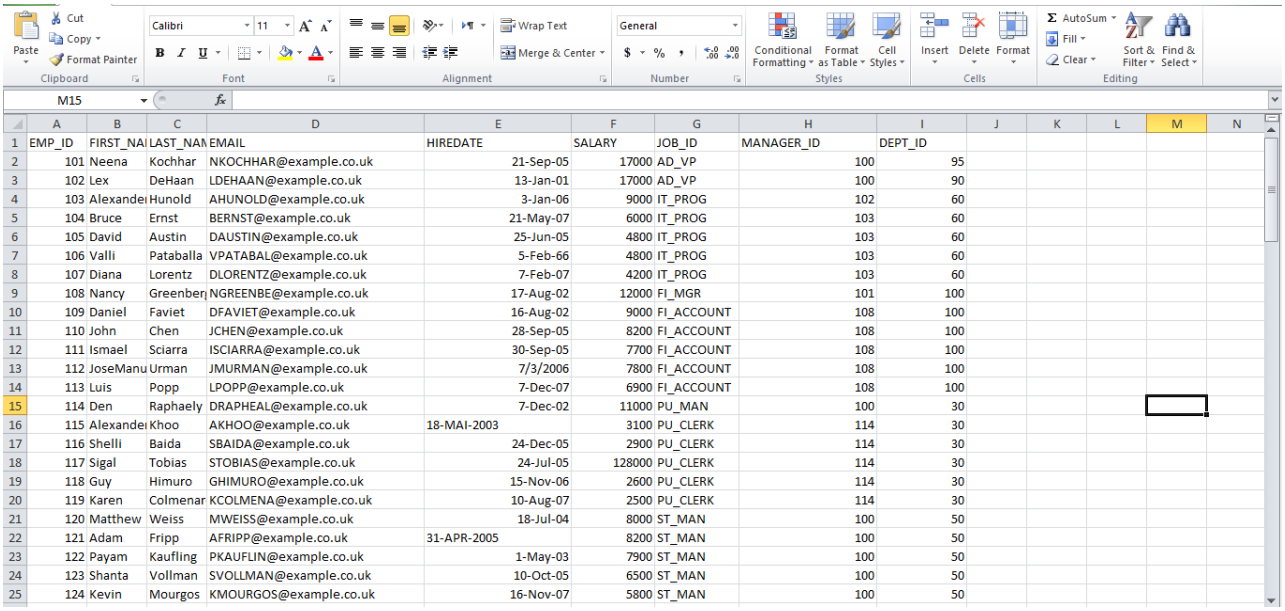
Delete the value cannot be replaced

Before:



EMP_ID	FIRST_NAI	LAST_NAN	EMAIL	HIREDATE	SALARY	JOB_ID	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	Alexande	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
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	JoseManu	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	Alexande	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	Colmenar	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
122	Payam	Kaufling	PKAUFLIN@example.co.uk	1-May-03	7900	ST_MAN	100	50
123	Shanta	Vollman	SVOLLMAN@example.co.uk	10-Oct-05	6500	ST_MAN	100	50

After

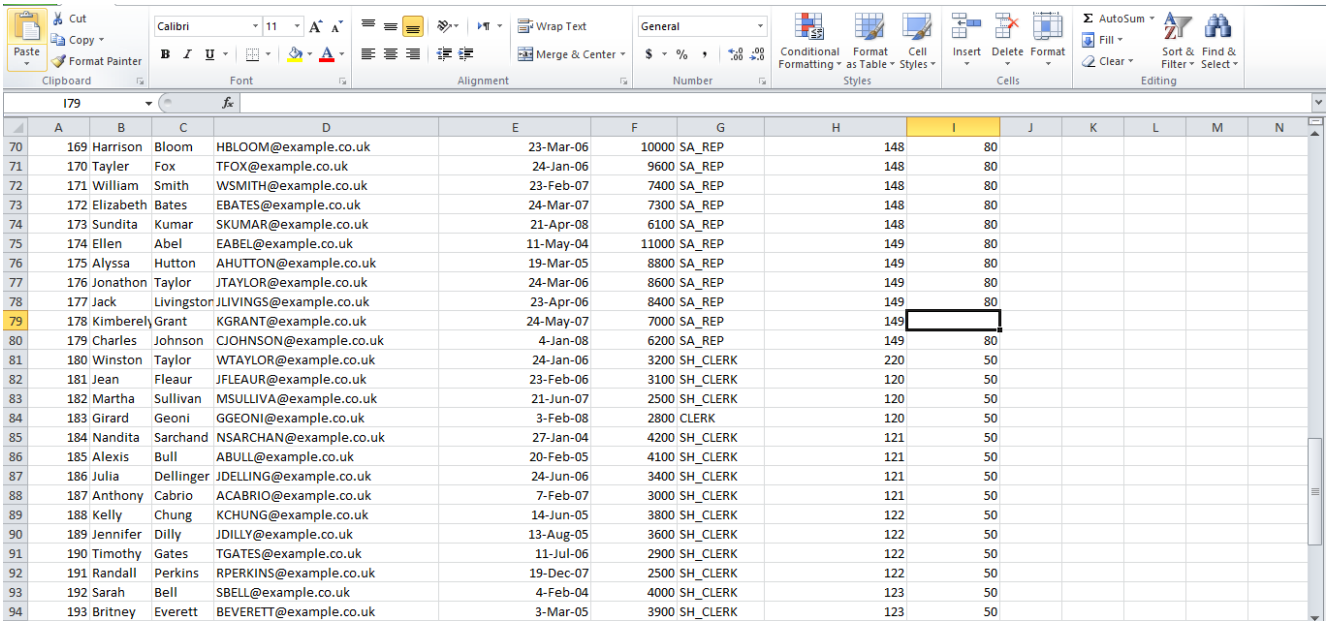


EMP_ID	FIRST_NAI	LAST_NAN	EMAIL	HIREDATE	SALARY	JOB_ID	MANAGER_ID	DEPT_ID
101	Neena	Kochhar	NKOCHHAR@example.co.uk	21-Sep-05	17000	AD_VP	100	95
102	Lex	DeHaan	LDEHAAN@example.co.uk	13-Jan-01	17000	AD_VP	100	90
103	Alexande	Hunold	AHUNOLD@example.co.uk	3-Jan-06	9000	IT_PROG	102	60
104	Bruce	Ernst	BERNST@example.co.uk	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
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	JoseManu	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	Alexande	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	Colmenar	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
122	Payam	Kaufling	PKAUFLIN@example.co.uk	1-May-03	7900	ST_MAN	100	50
123	Shanta	Vollman	SVOLLMAN@example.co.uk	10-Oct-05	6500	ST_MAN	100	50
124	Kevin	Mourgos	KMOURGOS@example.co.uk	16-Nov-07	5800	ST_MAN	100	50

From row 2 First_Name: steven is removed since it has no value in Manger_ID column.

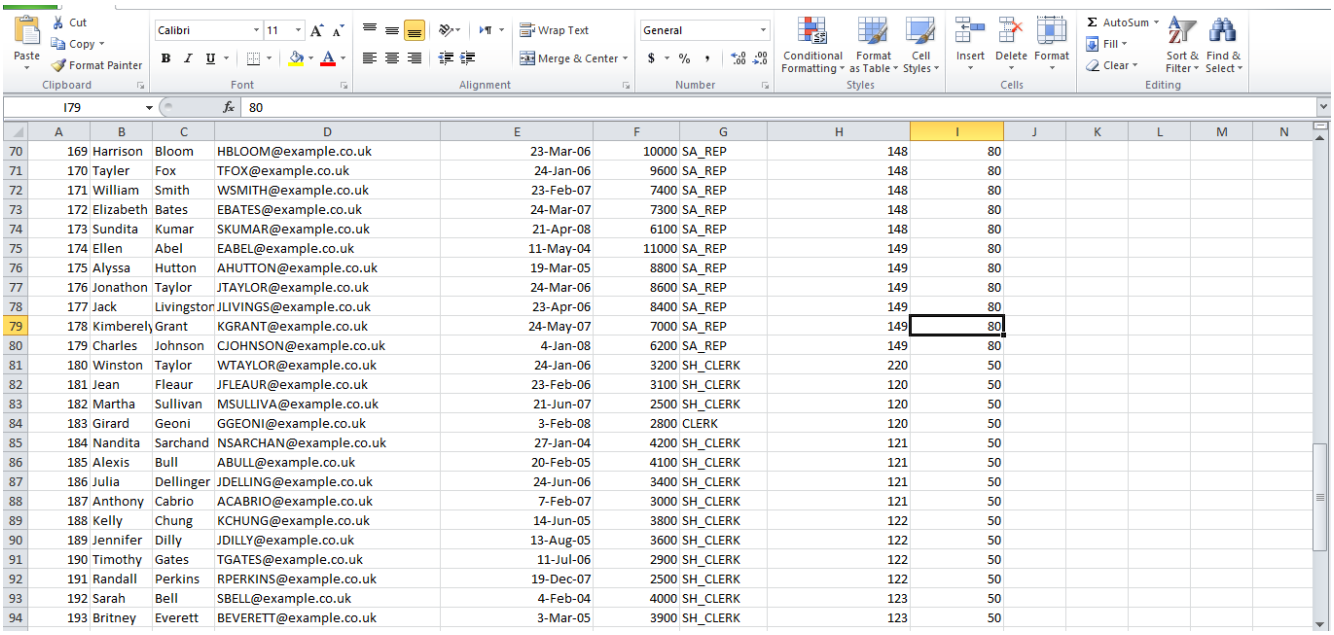
Putting Data in Empty Filed

Before:



	A	B	C	D	E	F	G	H	I
70	169	Harrison	Bloom	HBLOOM@example.co.uk	23-Mar-06	10000	SA_REP	148	80
71	170	Tayler	Fox	TFOX@example.co.uk	24-Jan-06	9600	SA_REP	148	80
72	171	William	Smith	WSMITH@example.co.uk	23-Feb-07	7400	SA_REP	148	80
73	172	Elizabeth	Bates	EBATES@example.co.uk	24-Mar-07	7300	SA_REP	148	80
74	173	Sundita	Kumar	SKUMAR@example.co.uk	21-Apr-08	6100	SA_REP	148	80
75	174	Ellen	Abel	EABEL@example.co.uk	11-May-04	11000	SA_REP	149	80
76	175	Alyssa	Hutton	AHUTTON@example.co.uk	19-Mar-05	8800	SA_REP	149	80
77	176	Jonathon	Taylor	JTAYLOR@example.co.uk	24-Mar-06	8600	SA_REP	149	80
78	177	Jack	Livingstor	JLIVINGS@example.co.uk	23-Apr-06	8400	SA_REP	149	80
79	178	Kimberely	Grant	KGRANT@example.co.uk	24-May-07	7000	SA_REP	149	
80	179	Charles	Johnson	CJOHNSON@example.co.uk	4-Jan-08	6200	SA_REP	149	80
81	180	Winston	Taylor	WTAYLOR@example.co.uk	24-Jan-06	3200	SH_CLERK	220	50
82	181	Jean	Fleaur	JFLEAUR@example.co.uk	23-Feb-06	3100	SH_CLERK	120	50
83	182	Martha	Sullivan	MSULLIVA@example.co.uk	21-Jun-07	2500	SH_CLERK	120	50
84	183	Girard	Geoni	GGEONI@example.co.uk	3-Feb-08	2800	CLERK	120	50
85	184	Nandita	Sarchand	NSARCHAN@example.co.uk	27-Jan-04	4200	SH_CLERK	121	50
86	185	Alexis	Bull	ABULL@example.co.uk	20-Feb-05	4100	SH_CLERK	121	50
87	186	Julia	Dellinger	JDELLING@example.co.uk	24-Jun-06	3400	SH_CLERK	121	50
88	187	Anthony	Cabrio	ACABRIO@example.co.uk	7-Feb-07	3000	SH_CLERK	121	50
89	188	Kelly	Chung	KCHUNG@example.co.uk	14-Jun-05	3800	SH_CLERK	122	50
90	189	Jennifer	Dilly	JDILLY@example.co.uk	13-Aug-05	3600	SH_CLERK	122	50
91	190	Timothy	Gates	TGATES@example.co.uk	11-Jul-06	2900	SH_CLERK	122	50
92	191	Randall	Perkins	RPERKINS@example.co.uk	19-Dec-07	2500	SH_CLERK	122	50
93	192	Sarah	Bell	SBELL@example.co.uk	4-Feb-04	4000	SH_CLERK	123	50
94	193	Britney	Everett	BEVERETT@example.co.uk	3-Mar-05	3900	SH_CLERK	123	50

After:

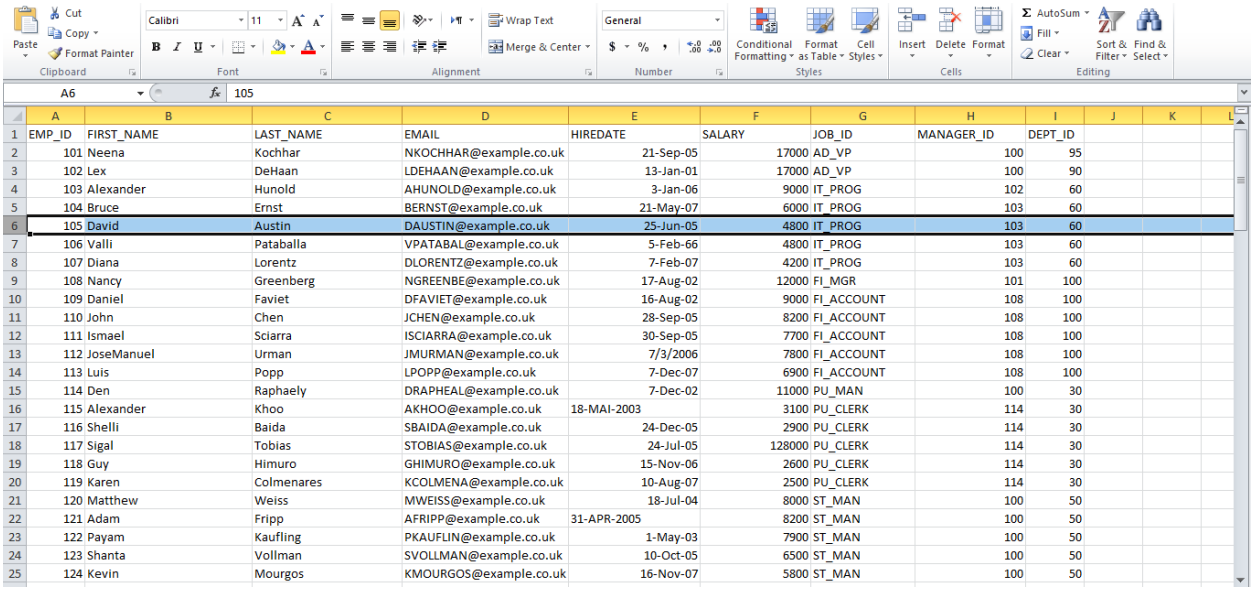


	A	B	C	D	E	F	G	H	I
70	169	Harrison	Bloom	HBLOOM@example.co.uk	23-Mar-06	10000	SA_REP	148	80
71	170	Tayler	Fox	TFOX@example.co.uk	24-Jan-06	9600	SA_REP	148	80
72	171	William	Smith	WSMITH@example.co.uk	23-Feb-07	7400	SA_REP	148	80
73	172	Elizabeth	Bates	EBATES@example.co.uk	24-Mar-07	7300	SA_REP	148	80
74	173	Sundita	Kumar	SKUMAR@example.co.uk	21-Apr-08	6100	SA_REP	148	80
75	174	Ellen	Abel	EABEL@example.co.uk	11-May-04	11000	SA_REP	149	80
76	175	Alyssa	Hutton	AHUTTON@example.co.uk	19-Mar-05	8800	SA_REP	149	80
77	176	Jonathon	Taylor	JTAYLOR@example.co.uk	24-Mar-06	8600	SA_REP	149	80
78	177	Jack	Livingstor	JLIVINGS@example.co.uk	23-Apr-06	8400	SA_REP	149	80
79	178	Kimberely	Grant	KGRANT@example.co.uk	24-May-07	7000	SA_REP	149	80
80	179	Charles	Johnson	CJOHNSON@example.co.uk	4-Jan-08	6200	SA_REP	149	80
81	180	Winston	Taylor	WTAYLOR@example.co.uk	24-Jan-06	3200	SH_CLERK	220	50
82	181	Jean	Fleaur	JFLEAUR@example.co.uk	23-Feb-06	3100	SH_CLERK	120	50
83	182	Martha	Sullivan	MSULLIVA@example.co.uk	21-Jun-07	2500	SH_CLERK	120	50
84	183	Girard	Geoni	GGEONI@example.co.uk	3-Feb-08	2800	CLERK	120	50
85	184	Nandita	Sarchand	NSARCHAN@example.co.uk	27-Jan-04	4200	SH_CLERK	121	50
86	185	Alexis	Bull	ABULL@example.co.uk	20-Feb-05	4100	SH_CLERK	121	50
87	186	Julia	Dellinger	JDELLING@example.co.uk	24-Jun-06	3400	SH_CLERK	121	50
88	187	Anthony	Cabrio	ACABRIO@example.co.uk	7-Feb-07	3000	SH_CLERK	121	50
89	188	Kelly	Chung	KCHUNG@example.co.uk	14-Jun-05	3800	SH_CLERK	122	50
90	189	Jennifer	Dilly	JDILLY@example.co.uk	13-Aug-05	3600	SH_CLERK	122	50
91	190	Timothy	Gates	TGATES@example.co.uk	11-Jul-06	2900	SH_CLERK	122	50
92	191	Randall	Perkins	RPERKINS@example.co.uk	19-Dec-07	2500	SH_CLERK	122	50
93	192	Sarah	Bell	SBELL@example.co.uk	4-Feb-04	4000	SH_CLERK	123	50
94	193	Britney	Everett	BEVERETT@example.co.uk	3-Mar-05	3900	SH_CLERK	123	50

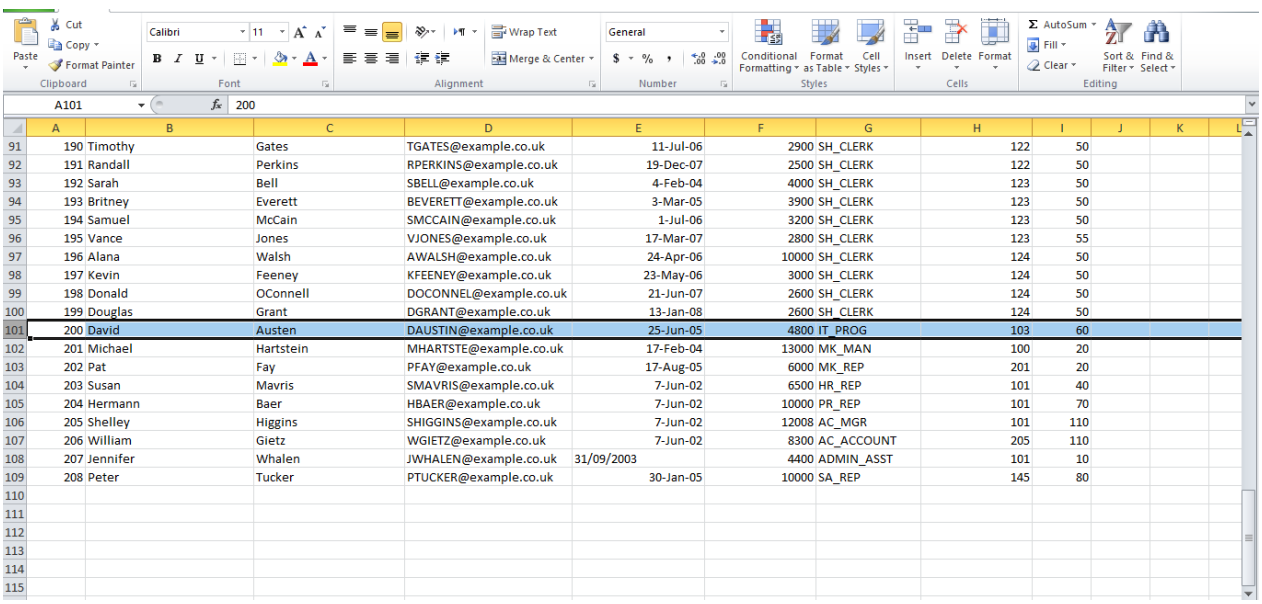
As seen in the row 79, Dept_ID column there is empty value, so in second figure, data is put in empty filed.

Duplicate Data

Before:



	A	B	C	D	E	F	G	H	I
1	EMP_ID	FIRST_NAME	LAST_NAME	EMAIL	HIREDATE	SALARY	JOB_ID	MANAGER_ID	DEPT_ID
2	101	Neena	Kochhar	NKOCHHAR@example.co.uk	21-Sep-05	17000	AD_VP	100	95
3	102	Lex	DeHaan	LDEHAAN@example.co.uk	13-Jan-01	17000	AD_VP	100	90
4	103	Alexander	Hunold	AHUNOLD@example.co.uk	3-Jan-06	9000	IT_PROG	102	60
5	104	Bruce	Ernst	BERNST@example.co.uk	21-May-07	6000	IT_PROG	103	60
6	105	David	Austin	DAUSTIN@example.co.uk	25-Jun-05	4800	IT_PROG	103	60
7	106	Valli	Pataballa	VPATABAL@example.co.uk	5-Feb-66	4800	IT_PROG	103	60
8	107	Diana	Lorentz	DLORENTZ@example.co.uk	7-Feb-07	4200	IT_PROG	103	60
9	108	Nancy	Greenberg	NGREENBE@example.co.uk	17-Aug-02	12000	FI_MGR	101	100
10	109	Daniel	Faviet	DFAVIET@example.co.uk	16-Aug-02	9000	FI_ACCOUNT	108	100
11	110	John	Chen	JCHEN@example.co.uk	28-Sep-05	8200	FI_ACCOUNT	108	100
12	111	Ismael	Sciarra	ISCIARRA@example.co.uk	30-Sep-05	7700	FI_ACCOUNT	108	100
13	112	JoseManuel	Urman	JMURMAN@example.co.uk	7/3/2006	7800	FI_ACCOUNT	108	100
14	113	Luis	Popp	LPOPP@example.co.uk	7-Dec-07	6900	FI_ACCOUNT	108	100
15	114	Den	Raphaely	DRAPHEAL@example.co.uk	7-Dec-02	11000	PU_MAN	100	30
16	115	Alexander	Khoo	AKHOO@example.co.uk	18-MAI-2003	3100	PU_CLERK	114	30
17	116	Shelli	Baida	SBAIDA@example.co.uk	24-Dec-05	2900	PU_CLERK	114	30
18	117	Sigal	Tobias	STOBIAS@example.co.uk	24-Jul-05	128000	PU_CLERK	114	30
19	118	Guy	Himuro	GHIMURO@example.co.uk	15-Nov-06	2600	PU_CLERK	114	30
20	119	Karen	Colmenares	KCOLMENA@example.co.uk	10-Aug-07	2500	PU_CLERK	114	30
21	120	Matthew	Weiss	MWEISS@example.co.uk	18-Jul-04	8000	ST_MAN	100	50
22	121	Adam	Fripp	AFRIPP@example.co.uk	31-APR-2005	8200	ST_MAN	100	50
23	122	Payam	Kaufling	PKAUFLIN@example.co.uk	1-May-03	7900	ST_MAN	100	50
24	123	Shanta	Vollman	SVOLLMAN@example.co.uk	10-Oct-05	6500	ST_MAN	100	50
25	124	Kevin	Mourgos	KMOURGOS@example.co.uk	16-Nov-07	5800	ST_MAN	100	50



	A	B	C	D	E	F	G	H	I
91	190	Timothy	Gates	TGATES@example.co.uk	11-Jul-06	2900	SH_CLERK	122	50
92	191	Randall	Perkins	RPERKINS@example.co.uk	19-Dec-07	2500	SH_CLERK	122	50
93	192	Sarah	Bell	SBELL@example.co.uk	4-Feb-04	4000	SH_CLERK	123	50
94	193	Britney	Everett	BEVERETT@example.co.uk	3-Mar-05	3900	SH_CLERK	123	50
95	194	Samuel	McCain	SMCCAIN@example.co.uk	1-Jul-06	3200	SH_CLERK	123	50
96	195	Vance	Jones	VJONES@example.co.uk	17-Mar-07	2800	SH_CLERK	123	55
97	196	Alana	Walsh	AWALSH@example.co.uk	24-Apr-06	10000	SH_CLERK	124	50
98	197	Kevin	Feeney	KFEENEY@example.co.uk	23-May-06	3000	SH_CLERK	124	50
99	198	Donald	OConnell	DOCONNEL@example.co.uk	21-Jun-07	2600	SH_CLERK	124	50
100	199	Douglas	Grant	DGRANT@example.co.uk	13-Jan-08	2600	SH_CLERK	124	50
101	200	David	Austen	DAUSTIN@example.co.uk	25-Jun-05	4800	IT_PROG	103	60
102	201	Michael	Hartstein	MHARTSTE@example.co.uk	17-Feb-04	13000	MK_MAN	100	20
103	202	Pat	Fay	PFAY@example.co.uk	17-Aug-05	6000	MK_REP	201	20
104	203	Susan	Mavris	SMAVRIS@example.co.uk	7-Jun-02	6500	HR_REP	101	40
105	204	Hermann	Baer	HBAER@example.co.uk	7-Jun-02	10000	PR_REP	101	70
106	205	Shelley	Higgins	SHIGGINS@example.co.uk	7-Jun-02	12008	AC_MGR	101	110
107	206	William	Gietz	WGIETZ@example.co.uk	7-Jun-02	8300	AC_ACCOUNT	205	110
108	207	Jennifer	Whalen	JWHALEN@example.co.uk	31/09/2003	4400	ADMIN_ASST	101	10
109	208	Peter	Tucker	PTUCKER@example.co.uk	30-Jan-05	10000	SA_REP	145	80

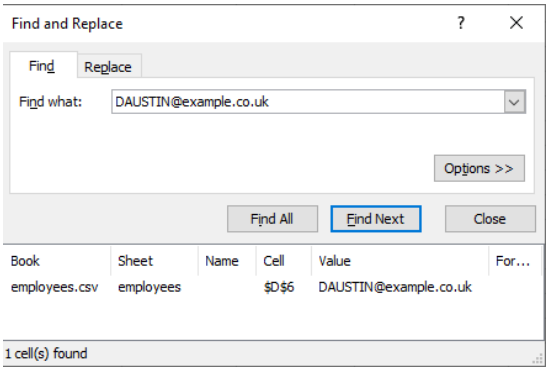
After:

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

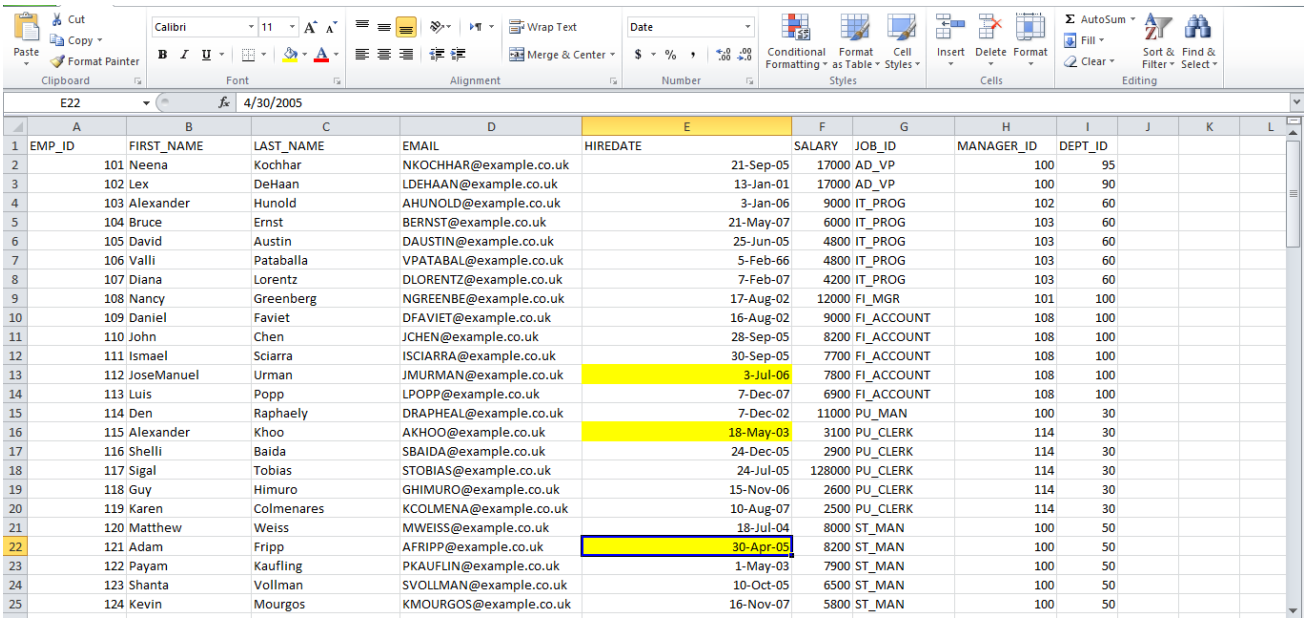
Inconsistent Data

Before:

EMP_ID	FIRST_NAME	LAST_NAME	EMAIL	HIREDATE	SALARY	JOB_ID	MANAGER_ID	DEPT_ID
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
122	Payam	Kaufling	PKAUFLIN@example.co.uk	1-May-03	7900	ST_MAN	100	50
123	Shanta	Vollman	SVOLLMAN@example.co.uk	10-Oct-05	6500	ST_MAN	100	50
124	Kevin	Mourgos	KMOURGOS@example.co.uk	16-Nov-07	5800	ST_MAN	100	50



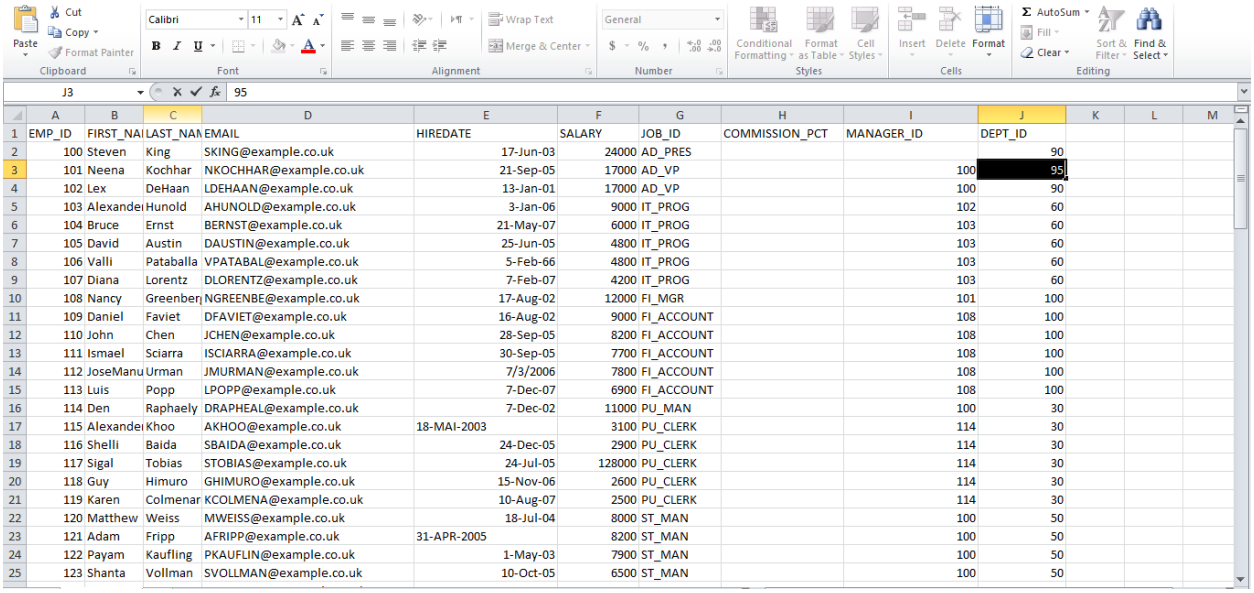
After:



EMP_ID	FIRST_NAME	LAST_NAME	EMAIL	HIREDATE	SALARY	JOB_ID	MANAGER_ID	DEPT_ID
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	3-Jul-06	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-May-03	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	30-Apr-05	8200	ST_MAN	100	50
122	Payam	Kaufling	PKAUFLIN@example.co.uk	1-May-03	7900	ST_MAN	100	50
123	Shanta	Vollman	SVOLLMAN@example.co.uk	10-Oct-05	6500	ST_MAN	100	50
124	Kevin	Mourgos	KMOURGOS@example.co.uk	16-Nov-07	5800	ST_MAN	100	50

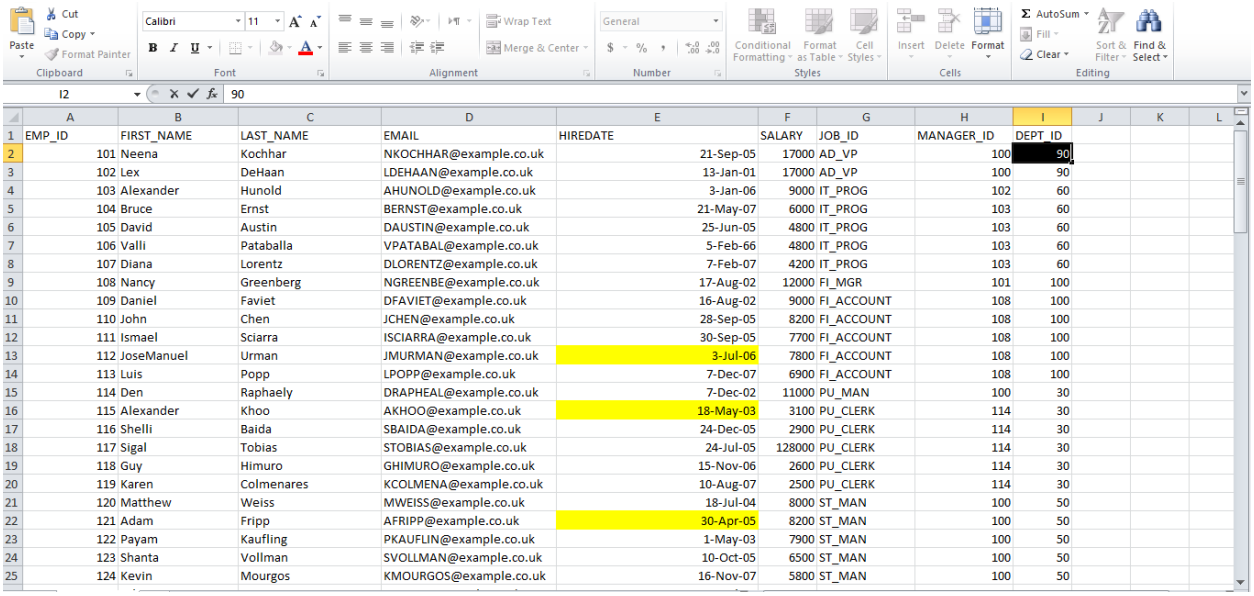
Incorrect Data:

Before:



EMP_ID	FIRST_NAI	LAST_NAI	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	Alexandei	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
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	JoseManu	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	Alexandei	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	Colmenar	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
122	Payam	Kaufling	PKAUFLIN@example.co.uk	1-May-03	7900	ST_MAN		100	50
123	Shanta	Vollman	SVOLLMAN@example.co.uk	10-Oct-05	6500	ST_MAN		100	50

After:

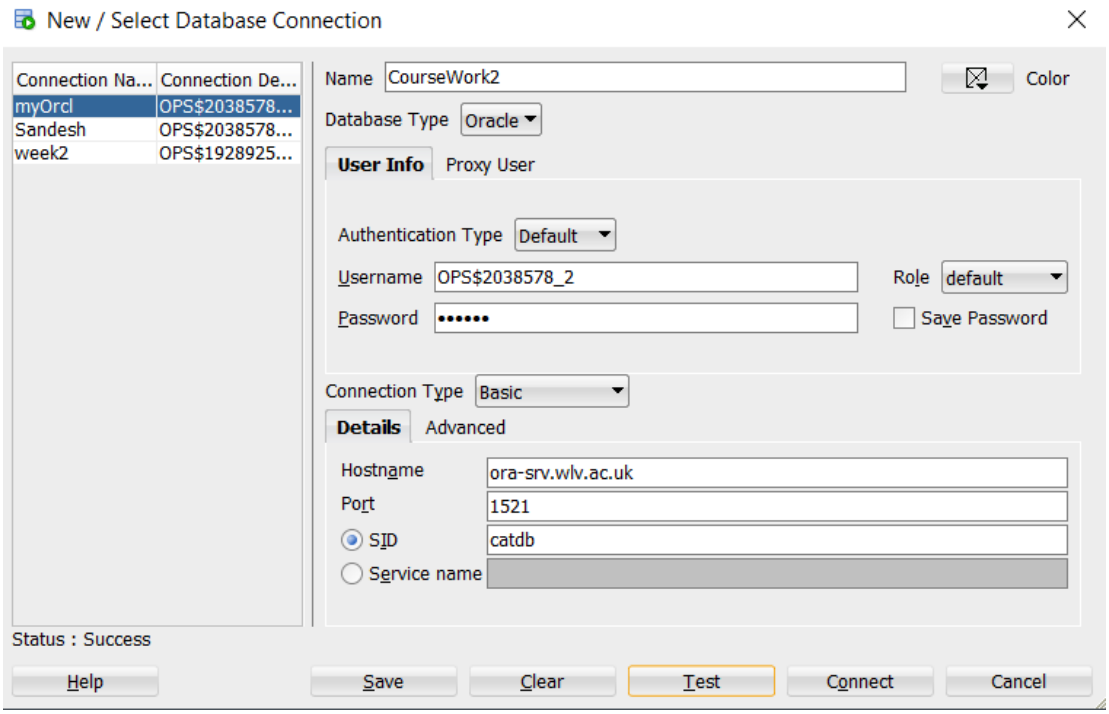


EMP_ID	FIRST_NAME	LAST_NAME	EMAIL	HIREDATE	SALARY	JOB_ID	MANAGER_ID	DEPT_ID
101	Neena	Kochhar	NKOCHHAR@example.co.uk	21-Sep-05	17000	AD_VP	100	90
102	Lex	DeHaan	LDEHAAN@example.co.uk	13-Jan-01	17000	AD_VP	100	90
103	Alexander	Hunold	AHUNOLD@example.co.uk	3-Jan-06	9000	IT_PROG	102	60
104	Bruce	Ernst	BERNST@example.co.uk	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	3-Jul-06	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-May-03	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	30-Apr-05	8200	ST_MAN	100	50
122	Payam	Kaufling	PKAUFLIN@example.co.uk	1-May-03	7900	ST_MAN	100	50
123	Shanta	Vollman	SVOLLMAN@example.co.uk	10-Oct-05	6500	ST_MAN	100	50
124	Kevin	Mourgos	KMOURGOS@example.co.uk	16-Nov-07	5800	ST_MAN	100	50

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.

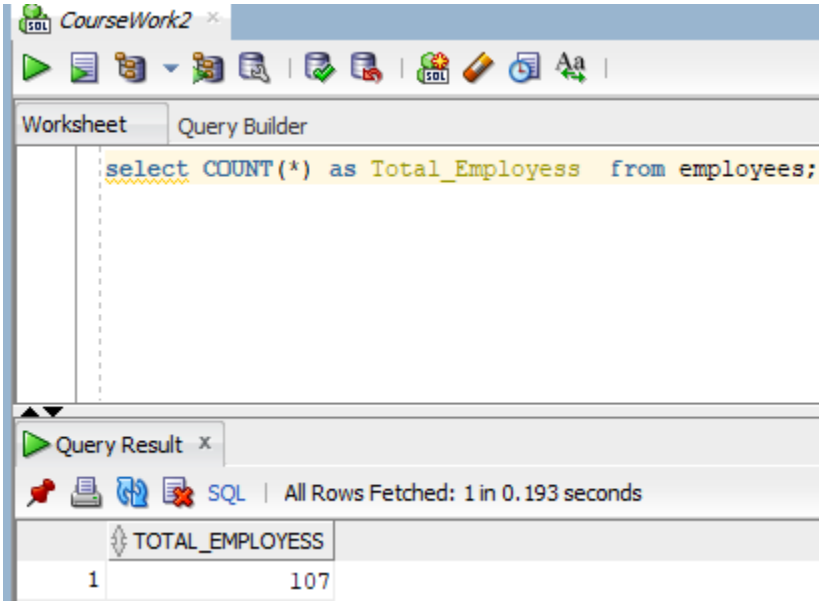
Importing

Creating new connection

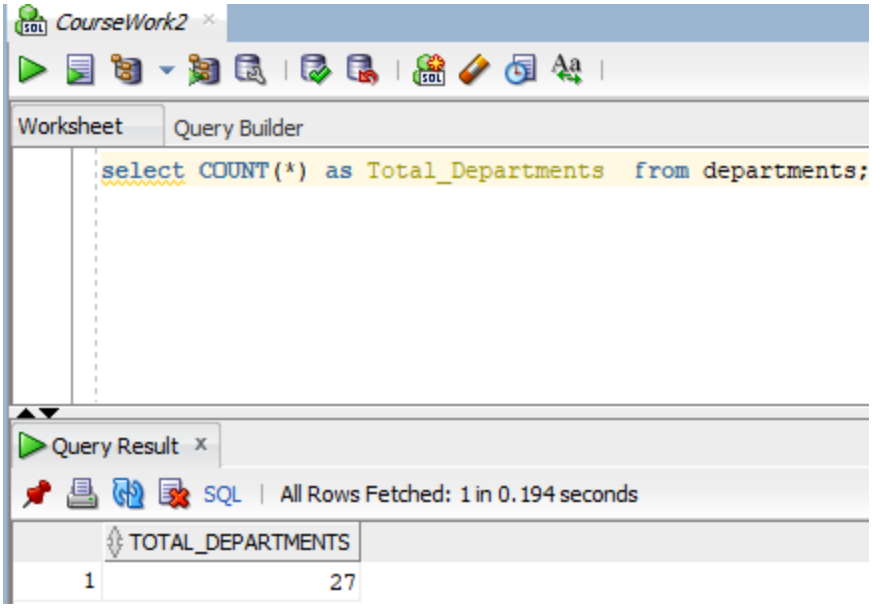


SQL queries

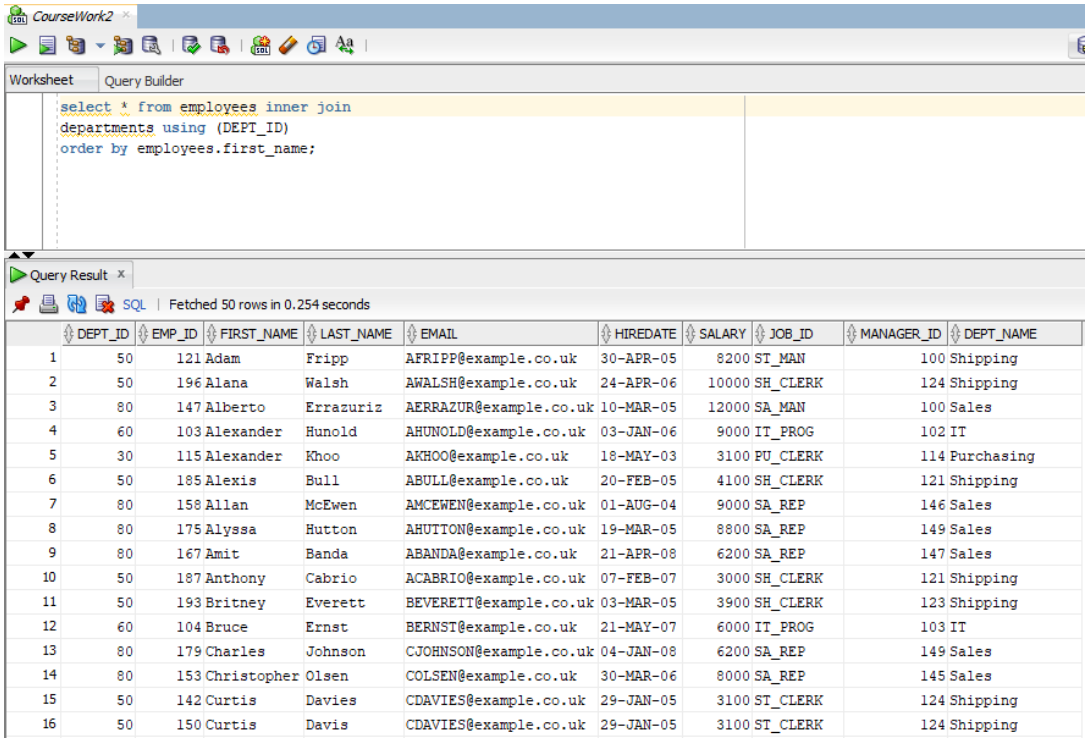
Counting total number of emp:



Counting number of total dept



Join Emp and Dep table:



	DEPT_ID	EMP_ID	FIRST_NAME	LAST_NAME	EMAIL	HIREDATE	SALARY	JOB_ID	MANAGER_ID	DEPT_NAME
1	50	121	Adam	Fripp	AFRIPP@example.co.uk	30-APR-05	8200	ST_MAN	100	Shipping
2	50	196	Alana	Walsh	AWALSH@example.co.uk	24-APR-06	10000	SH_CLERK	124	Shipping
3	80	147	Alberto	Errazuriz	AERRAZUR@example.co.uk	10-MAR-05	12000	SA_MAN	100	Sales
4	60	103	Alexander	Hunold	AHUNOLD@example.co.uk	03-JAN-06	9000	IT_PROG	102	IT
5	30	115	Alexander	Khoo	AKHOO@example.co.uk	18-MAY-03	3100	PU_CLERK	114	Purchasing
6	50	185	Alexis	Bull	ABULL@example.co.uk	20-FEB-05	4100	SH_CLERK	121	Shipping
7	80	158	Allan	McEwen	AMCEWEN@example.co.uk	01-AUG-04	9000	SA_REP	146	Sales
8	80	175	Alyssa	Hutton	AHUTTON@example.co.uk	19-MAR-05	8800	SA_REP	149	Sales
9	80	167	Amit	Banda	ABANDA@example.co.uk	21-APR-08	6200	SA_REP	147	Sales
10	50	187	Anthony	Cabrio	ACABRIO@example.co.uk	07-FEB-07	3000	SH_CLERK	121	Shipping
11	50	193	Britney	Everett	BEVERETT@example.co.uk	03-MAR-05	3900	SH_CLERK	123	Shipping
12	60	104	Bruce	Ernst	BERNST@example.co.uk	21-MAY-07	6000	IT_PROG	103	IT
13	80	179	Charles	Johnson	CJOHNSON@example.co.uk	04-JAN-08	6200	SA_REP	149	Sales
14	80	153	Christopher	Olsen	COLSEN@example.co.uk	30-MAR-06	8000	SA_REP	145	Sales
15	50	142	Curtis	Davies	CDAVIES@example.co.uk	29-JAN-05	3100	ST_CLERK	124	Shipping
16	50	150	Curtis	Davis	CDAVIES@example.co.uk	29-JAN-05	3100	ST_CLERK	124	Shipping

Cube Command:

CourseWork2

Worksheet | Query Builder

```
select  first_name,  dept_id, round(sum(salary),2) as wages
from employees
group by cube( first_name,  dept_id)
order by first_name;
```

Query Result

SQL | Fetched 50 rows in 0.228 seconds

	FIRST_NAME	DEPT_ID	WAGES
1	Adam	50	8200
2	Adam	(null)	8200
3	Alana	50	10000
4	Alana	(null)	10000
5	Alberto	80	12000
6	Alberto	(null)	12000
7	Alexander	30	3100
8	Alexander	60	9000
9	Alexander	(null)	12100
10	Alexis	50	4100
11	Alexis	(null)	4100
12	Allan	80	9000
13	Allan	(null)	9000
14	Alyssa	80	8800
15	Alyssa	(null)	8800
16	Amit	80	6200
17	Amit	(null)	6200
18	Anthony	50	3000

Rollup command

CourseWork2

Worksheet | Query Builder

```
select emp_id, first_name, last_name, email, hiredate, job_id, dept_id, round(sum(salary),2) as wages
from employees
group by rollup(emp_id, first_name, last_name, email, hiredate, job_id, dept_id, salary)
order by dept_id;
```

Query Result

SQL | Fetched 50 rows in 0.352 seconds

	EMP_ID	FIRST_NAME	LAST_NAME	EMAIL	HIREDATE	JOB_ID	DEPT_ID	WAGES
1	207	Jennifer	Whalen	JWHALEN@example.co.uk	30-SEP-03	ADMIN_ASST	10	4400
2	207	Jennifer	Whalen	JWHALEN@example.co.uk	30-SEP-03	ADMIN_ASST	10	4400
3	201	Michael	Hartstein	MHARTSTE@example.co.uk	17-FEB-04	MK_MAN	20	13000
4	201	Michael	Hartstein	MHARTSTE@example.co.uk	17-FEB-04	MK_MAN	20	13000
5	202	Pat	Fay	PFAY@example.co.uk	17-AUG-05	MK_REP	20	6000
6	202	Pat	Fay	PFAY@example.co.uk	17-AUG-05	MK_REP	20	6000
7	114	Den	Raphaely	DRAPHEAL@example.co.uk	07-DEC-02	PU_MAN	30	11000
8	119	Karen	Colmenares	KCOLMENA@example.co.uk	10-AUG-07	PU_CLERK	30	2500
9	114	Den	Raphaely	DRAPHEAL@example.co.uk	07-DEC-02	PU_MAN	30	11000
10	119	Karen	Colmenares	KCOLMENA@example.co.uk	10-AUG-07	PU_CLERK	30	2500
11	115	Alexander	Khoo	AKHOO@example.co.uk	18-MAY-03	PU_CLERK	30	3100
12	115	Alexander	Khoo	AKHOO@example.co.uk	18-MAY-03	PU_CLERK	30	3100
13	116	Shelli	Baida	SBAIDA@example.co.uk	24-DEC-05	PU_CLERK	30	2900
14	116	Shelli	Baida	SBAIDA@example.co.uk	24-DEC-05	PU_CLERK	30	2900
15	117	Sigal	Tobias	STOBIAS@example.co.uk	24-JUL-05	PU_CLERK	30	128000
16	117	Sigal	Tobias	STOBIAS@example.co.uk	24-JUL-05	PU_CLERK	30	128000
17	118	Guy	Himuro	GHIMURO@example.co.uk	15-NOV-06	PU_CLERK	30	2600
18	118	Guy	Himuro	GHIMURO@example.co.uk	15-NOV-06	PU_CLERK	30	2600

Visualization

Importing required libraries

```
[2]
import pandas as pd
import numpy as np
import matplotlib.pyplot as mp
```

Importing files from drive and reading the files using pandas

```
[33]
emp = pd.read_csv('/content/drive/MyDrive/Colab Notebooks/BigDataCourseWork/IndividualDataSet/employees.csv')
dep = pd.read_csv('/content/drive/MyDrive/Colab Notebooks/BigDataCourseWork/IndividualDataSet/departments.csv')
```

Showing the first 10 data of dept

```
dep.head(10)
```

	DEPT_ID	DEPT_NAME
0	10	Administration
1	20	Marketing
2	30	Purchasing
3	40	HumanResources
4	50	Shipping
5	60	IT
6	70	PublicRelations
7	80	Sales
8	90	Executive
9	100	Finance

Showing the first 10 data of emp

```
emp.head(10)
--NORMAL--
```

	EMP_ID	FIRST_NAME	LAST_NAME	EMAIL	HIREDATE	SALARY	JOB_ID	MANAGER_ID	DEPT_ID
0	101	Neena	Kochhar	NKOCHHAR@example.co.uk	21-Sep-05	17000	AD_VP	100	90
1	102	Lex	DeHaan	LDEHAAN@example.co.uk	13-Jan-01	17000	AD_VP	100	90
2	103	Alexander	Hunold	AHUNOLD@example.co.uk	3-Jan-06	9000	IT_PROG	102	60
3	104	Bruce	Ernst	BERNST@example.co.uk	21-May-07	6000	IT_PROG	103	60
4	105	David	Austin	DAUSTIN@example.co.uk	25-Jun-05	4800	IT_PROG	103	60
5	106	Valli	Pataballa	VPATABAL@example.co.uk	5-Feb-66	4800	IT_PROG	103	60
6	107	Diana	Lorentz	DLORENTZ@example.co.uk	7-Feb-07	4200	IT_PROG	103	60
7	108	Nancy	Greenberg	NGREENBE@example.co.uk	17-Aug-02	12000	FI_MGR	101	100
8	109	Daniel	Faviet	DFAVIET@example.co.uk	16-Aug-02	9000	FI_ACCOUNT	108	100
9	110	John	Chen	JCHEN@example.co.uk	28-Sep-05	8200	FI_ACCOUNT	108	100

Based on dept_id, emp and dept table are inner joined

```
[36]  join = pd.merge(emp,dep, on='DEPT_ID', how='inner')
```

Displaying the first 10 data result after joining

```
[37]  join.head(10)
```

	EMP_ID	FIRST_NAME	LAST_NAME	EMAIL	HIREDATE	SALARY	JOB_ID	MANAGER_ID	DEPT_ID	DEPT_NAME
0	101	Neena	Kochhar	NKOCHHAR@example.co.uk	21-Sep-05	17000	AD_VP	100	90	Executive
1	102	Lex	DeHaan	LDEHAAN@example.co.uk	13-Jan-01	17000	AD_VP	100	90	Executive
2	103	Alexander	Hunold	AHUNOLD@example.co.uk	3-Jan-06	9000	IT_PROG	102	60	IT
3	104	Bruce	Ernst	BERNST@example.co.uk	21-May-07	6000	IT_PROG	103	60	IT
4	105	David	Austin	DAUSTIN@example.co.uk	25-Jun-05	4800	IT_PROG	103	60	IT
5	106	Valli	Pataballa	VPATABAL@example.co.uk	5-Feb-66	4800	IT_PROG	103	60	IT
6	107	Diana	Lorentz	DLORENTZ@example.co.uk	7-Feb-07	4200	IT_PROG	103	60	IT
7	108	Nancy	Greenberg	NGREENBE@example.co.uk	17-Aug-02	12000	FI_MGR	101	100	Finance
8	109	Daniel	Faviet	DFAVIET@example.co.uk	16-Aug-02	9000	FI_ACCOUNT	108	100	Finance
9	110	John	Chen	JCHEN@example.co.uk	28-Sep-05	8200	FI_ACCOUNT	108	100	Finance

Summing the salary of emp and grouping by the dept_name and displaying the result

```
grp = join.groupby(['DEPT_NAME'])['SALARY'].sum()
grp
```

```
--INSERT--
```

```
DEPT_NAME
Accounting          20308
Administration       4400
Executive           34000
Finance             51600
HumanResources       6500
IT                  28800
Marketing           19000
PublicRelations     10000
Purchasing         150100
Sales              301000
Shipping           163600
Name: SALARY, dtype: int64
```

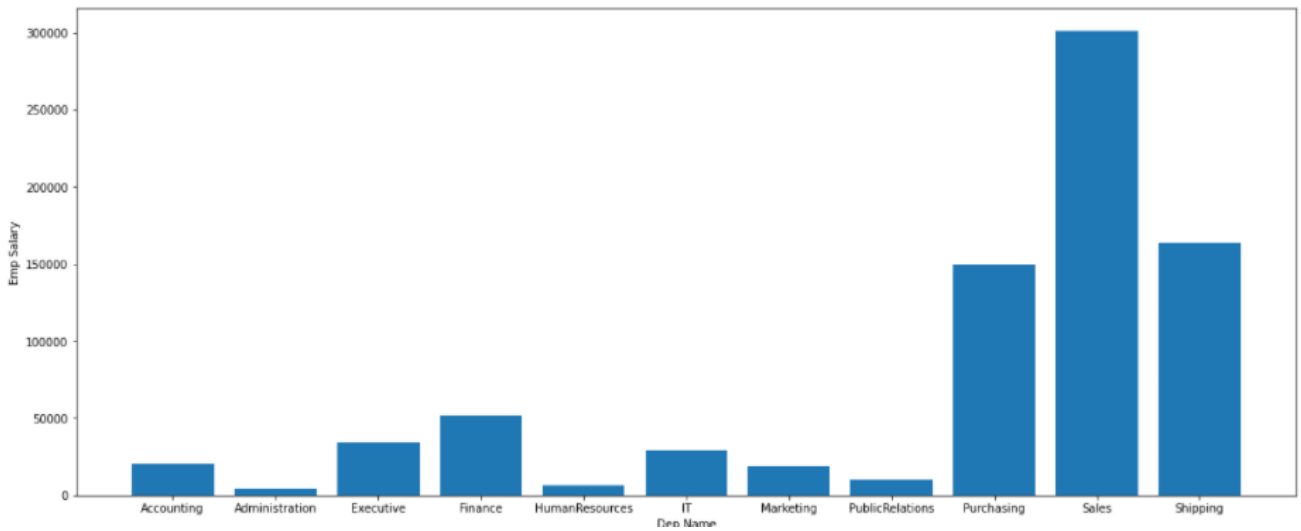
Changing the grouping result into python dict

```
[40]  hp = grp.to_dict()
      keys = list(hp.keys())
      values = list(hp.values())
```

Visualizing the above data in bar diagram and line diagram respectively

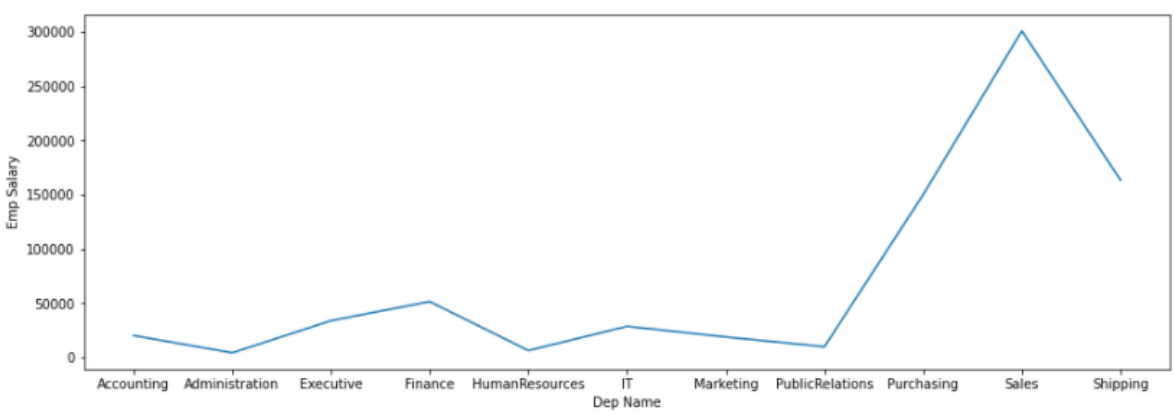
```
dell = mp.figure(figsize=(15,6))
op = dell.add_axes([0,0,1,1])
op.bar(keys, values)
mp.xlabel('Dep Name')
mp.ylabel('Emp Salary')
dell.show()
```

--INSERT--



[42]

```
dell = mp.figure(figsize=(15,5))
mp.plot(keys, values)
mp.xlabel('Dep Name')
mp.ylabel('Emp Salary')
mp.show()
```



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].