

# Hands-on Lab 5: Cleaning Data

---

**Estimated time needed:** 45 minutes

In this lab, first you will learn how to deal with inaccurate data, how to remove empty rows, and how to remove duplicated data. Next, you will learn how to change the case of text, how to change date formatting, and how to trim whitespace from data. Finally, you will learn how to use the Flash Fill feature and functions in Excel to help clean data.

## Software Used in this Lab

---

The instruction videos in this course use the full Excel Desktop version as this has all the available product features, but for the hands-on labs we will be using the free 'Excel for the web' version as this is available to everyone.

Although you can use the Excel Desktop software if you have access to this version, it is recommended that you use Excel for the web for the hands-on labs as the lab instructions specifically refer to this version, and there are some small differences in the interface and available features.

## Dataset Used in this Lab

---

The dataset used in this lab comes from the following source: <https://dataplatform.cloud.ibm.com/exchange/public/entry/view/f8ccaf607372882403a37d9019b3abf4>. This dataset is published by **IBM**, and includes fictitious customer demographics and sales data.

We are using a modified subset of that dataset for the lab, so to follow the lab instructions successfully please use the dataset provided with the lab, rather than the dataset from the original source.

## Objectives

---

After completing this lab, you will be able to:

- Understand how to deal with irrelevant or inaccurate data
- Remove empty rows and duplicated data
- Change text case and date formatting
- Trim whitespaces from data
- Use Flash Fill and functions to clean data

## Exercise 1: Removing Duplicated, Irrelevant or Inaccurate Data

---

In this exercise, you will learn how to deal with inaccurate data, how to remove empty rows, and how to remove duplicated data.

### Task A: Check spelling

---

1. Download the file **Customer\_demographics\_and\_sales\_Lab5.xlsx**. Upload and open it using Excel for the web.
2. Select column **L (CREDITCARD\_TYPE)**, then click **Review** tab, and select **Spelling**.
3. Click the correct suggestion to change the spelling.
  - **Note:** Don't change 'jcb' spelling when doing the spell check. We will need 'jcb' for the Exercise 1 Task D.
4. Close the **Spelling** pane.

K	L
NUMBER	CREDITCARD_TYPE
1-8539	Master Card
1-8539	Master Card
1-8539	Master Card
173271	VISA
5-4321	American Expres
5-4321	American Expres
5-4321	American Expres
5-4321	American Expres
5-4321	American Expres
5-4321	American Expres
5-4321	American Expres
2037	Diners Club
2037	Diners Club
2037	Diners Club
1865	VISA
1865	VISA
4-7595	Diners Club

Spelling

×

>

Not in Dictionary

American Expres

Suggestions

Express

Expires

Expreso

Ignore

Ignore All

## Task B: Remove empty rows

1. Press **CTRL+HOME**, then press **CTRL+SHIFT+END** to select the whole datasheet.
2. On the **Data** tab, click **Filter**.
3. Press **CTRL+HOME**, click the **filter arrow** in the **CUST\_NAME** column, and then click **Filter**.
4. Click the **Select All** checkbox to deselect all of them. Then select just **Blanks**, then **OK**.
5. Select **first row**, then press **CTRL+SHIFT+END** to select all rows.
6. Right-click the selected rows and then click **Delete Rows**.
7. Finally, on the **Data** tab, click **Clear**, then click **Filter**.

File

Home

Insert

Formulas

Data

Review

View

Help

Tell me what you want to do

Refresh Selected Connection

Refresh All Connections

Stocks

Geography

Sort Ascending

Sort Descending

Custom Sort

Filter

Clear

Reapply

A71

fx

A

B

C

D

1

CUSTNAME

GenderCod

ADDRESS1

CITY

1048558

1048559

1048560

1048561

1048562

1048563

1048564

1048565

1048566

1048567

1048568

1048569

1048570

1048571

1048572

1048573

1048574

1048575

1048576

1048577

1048578

1048579

1048580

1048581

1048582

1048583

1048584

1048585

1048586

1048587

1048588

1048589

1048590

1048591

1048592

1048593

1048594

1048595

1048596

1048597

1048598

1048599

1048600

1048601

1048602

1048603

1048604

1048605

1048606

1048607

1048608

1048609

1048610

1048611

1048612

1048613

1048614

1048615

1048616

1048617

1048618

1048619

1048620

1048621

1048622

1048623

1048624

1048625

1048626

1048627

1048628

1048629

1048630

1048631

1048632

1048633

1048634

1048635

1048636

1048637

1048638

1048639

1048640

1048641

1048642

1048643

1048644

1048645

1048646

1048647

1048648

1048649

1048650

1048651

1048652

1048653

1048654

1048655

1048656

1048657

1048658

1048659

1048660

1048661

1048662

1048663

1048664

1048665

1048666

1048667

1048668

1048669

1048670

1048671

1048672

1048673

1048674

1048675

1048676

1048677

1048678

1048679

1048680

1048681

1048682

1048683

1048684

1048685

1048686

1048687

1048688

1048689

1048690

1048691

1048692

1048693

1048694

1048695

1048696

1048697

1048698

1048699

1048700

1048701

1048702

1048703

1048704

1048705

1048706

1048707

1048708

1048709

1048710

1048711

1048712

1048713

1048714

1048715

1048716

1048717

1048718

1048719

1048720

1048721

1048722

1048723

1048724

1048725

1048726

1048727

1048728

1048729

1048730

1048731

1048732

1048733

1048734

1048735

1048736

1048737

1048738

1048739

1048740

1048741

1048742

1048743

1048744

1048745

1048746

1048747

1048748

1048749

1048750

1048751

1048752

1048753

1048754

1048755

1048756

1048757

1048758

1048759

1048760

1048761

1048762

1048763

1048764

1048765

1048766

1048767

1048768

1048769

1048770

1048771

1048772

1048773

1048774

1048775

1048776

1048777

1048778

1048779

1048780

1048781

1048782

1048783

1048784

1048785

1048786

1048787

1048788

1048789

1048790

1048791

1048792

1048793

1048794

1048795

1048796

1048797

1048798

1048799

1048800

1048801

1048802

1048803

1048804

1048805

1048806

1048807

1048808

1048809

1048810

1048811

1048812

1048813

1048814

1048815

1048816

1048817

1048818

1048819

1048820

1048821

1048822

1048823

1048824

1048825

1048826

1048827

1048828

1048829

1048830

1048831

1048832

1048833

1048834

1048835

1048836

1048837

1048838

1048839

1048840

1048841

1048842

1048843

1048844

1048845

1048846

1048847

1048848

1048849

1048850

1048851

1048852

1048853

1048854

1048855

1048856

1048857

1048858

1048859

1048860

1048861

1048862

1048863

1048864

1048865

1048866

1048867

1048868

1048869

1048870

1048871

1048872

1048873

1048874

1048875

1048876

1048877

1048878

1048879

1048880

1048881

1048882

1048883

1048884

1048885

1048886

1048887

1048888

1048889

1048890

1048891

1048892

1048893

1048894

1048895

1048896

1048897

1048898

1048899

1048900

1048901

1048902

1048903

1048904

1048905

1048906

1048907

1048908

1048909

1048910

1048911

1048912

1048913

1048914

1048915

1048916

1048917

1048918

1048919

1048920

1048921

1048922

1048923

1048924

1048925

1048926

1048927

1048928

1048929

1048930

1048931

1048932

1048933

1048934

1048935

1048936

1048937

1048938

1048939

1048940

1048941

1048942

1048943

1048944

1048945

1048946

1048947

1048948

1048949

1048950

1048951

1048952

1048953

1048954

1048955

1048956

1048957

1048958

1048959

1048960

1048961

1048962

1048963

1048964

1048965

1048966

1048967

1048968

1048969

1048970

1048971

1048972

1048973

1048974

1048975

1048976

1048977

1048978

1048979

1048980

1048981

1048982

1048983

1048984

1048985

1048986

1048987

1048988

1048989

1048990

1048991

1048992

1048993

1048994

1048995

1048996

1048997

1048998

1048999

1049000

Cut

Copy

Paste Options

Insert Rows

Delete Rows

Clear Contents

Row Height...

Hide Rows

Unhide Rows

Number Format...

## Task C: Remove duplicate rows

1. Select Column **T (ORDER\_ID)** since ORDER\_ID values are unique.
2. On the **Home** tab, click **Conditional Formatting> Highlight Cells Rules> Duplicate Values**, and then click **OK**.
3. Select the whole datasheet (**CTRL+SHIFT+END**)
4. On the **Data** tab, click **Remove Duplicates**.
5. In the Remove Duplicates dialog box, ensure that **Select all columns** is checked and that **My data has headers** is also checked, then click **OK**.
6. In the pop-up box informing you how many duplicate values were found and removed, click **OK**.

Excel

×

i

5 duplicate values found and removed; 194 unique values remain.

[Give Feedback](#)

OK

## Task D: Use Find & Replace to correct misspelling

1. On the **Home** tab, click **Find & Select**.
2. Click **Find**. In Find what, type **jcb**, and click **Find All**.
3. Click **Replace**.
4. In Replace with, type **JCB**, click **Replace All**, and then click the **Close** icon.
5. On the **Home** tab, click **Conditional Formatting > Clear Rules > Clear Rules from Entire Sheet**.

Find

Find what:  
jcb

Search Options

Within: Sheet Direction: Down

☒ Match case  
☒ Match entire cell contents

Results found (25)

Sheet	Cell	Value
Sheet1	L45	jcb
Sheet1	L46	jcb
Sheet1	L47	jcb
Sheet1	L49	jcb
Sheet1	L69	jcb
Sheet1	L70	jcb

Find Next Find All Replace...

## Exercise 2: Dealing with Inconsistencies in Data

In this exercise, you will learn how to change the case of text, how to change date formatting, and how to trim whitespace from data.

### Task A: Use the PROPER function to change text from upper case to proper case

1. Select row **2**, then right-click it and choose **Insert Rows**.
2. In cell **A2**, type **=PROPER(A1)** and press **Enter**.
3. Hover over the bottom-right corner of cell **A2**, and drag the **Fill Handle** across to the last column.
  - If dragging across is too difficult with the mouse, then select the cells in the row 2 using **SHIFT+RIGHT ARROW**, then press **F2** to put the cursor focus back in cell **A2**, then hold **CTRL** while you press **Enter**.
4. Select row **2**, then press **CTRL+C**.
5. Select row **1**, Right-click and choose **Paste Options>Values**.
6. Select row **2**, right-click it and choose **Delete Rows**.

### Task B: Use the UPPER function to change text from proper case to upper case

1. Select column **AG (Generation)**. Then right-click and choose **Insert Columns**. In cell **AG1**, type **Generation**.
2. In cell **AG2**, type **=UPPER(AH2)** and press **Enter**.
3. Hover over the bottom-right corner of cell **AG2** and double-click the **Fill Handle**.
4. Select column **AG**, then press **CTRL+C**.
5. Select column **AH**, right-click and choose **Paste Options>Values**.
6. Select column **AG**, right-click it and choose **Delete Columns**.

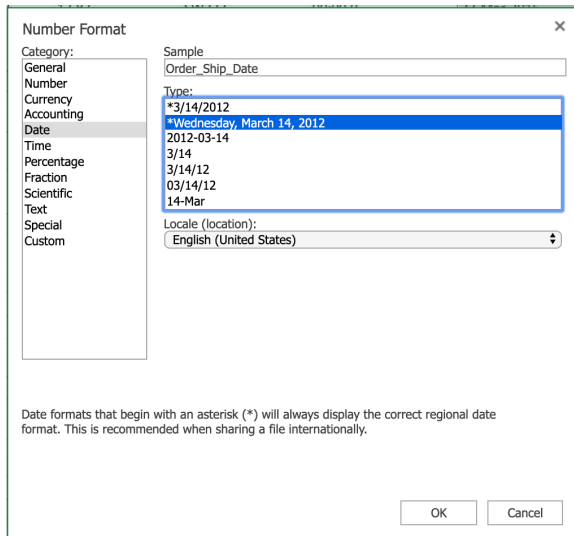
### Task C: Use the LOWER function to change text from proper case to lower case

1. Select column **AC (T\_Type)**. Then right-click and choose **Insert Columns**. In cell **AC1**, type **T\_Type**.

2. In cell **AC2**, type **=LOWER(AD2)** and press **Enter**.
3. Hover over the bottom-right corner of cell **AC2** and double-click the **Fill Handle**.
4. Select column **AC**, then press **CTRL+C**.
5. Select column **AD**, right-click and choose **Paste Options>Values**.
6. Select column **AC**, right-click it and choose **Delete Columns**.

## Task D: Change date formatting

1. Select column **Z (Order\_Ship\_Date)**.
2. On the **Home** tab, in the **Number** group click **Number Format> More Number Formats**.
3. In the Category list, select **Date**.
4. In the **Format Cells** box, under **Locale**, select **English (United States)**.
5. Under **Type**, select **Wednesday, March 14, 2012** and click **OK**.



## Task E: Use Find & Replace to trim whitespace

1. Click **CTRL+HOME**.
2. Select all the data using **CTRL+SHIFT+END**.
3. On the **Home** tab, click **Find & Select**, then **Replace**.
4. In Find what, type **2 spaces**. In Replace with, type **1 space**.
5. Click **Find All**, then click **Replace All**.
6. Click the **Close** icon.

## Exercise 3: More Excel Features for Cleaning Data

In this exercise, you will learn how to use the Flash Fill feature and functions in Excel to help clean data.

### Task A: Use the Flash Fill feature to clean data:

1. Select column **A (Cust\_Name)**, right-click and choose **Insert Columns**.
2. In cell **A1** type **Customer\_Name** and press **Enter**.
3. In cell **A2**, type **Mr. Allen Perl** and press **Enter**.
4. Select column **A (Customer\_Name)**, on the **Data** tab, click **Flash Fill**.
5. Click **Undo** to undo this step.

If you are using the desktop version of Excel, you could use the 'Text to Columns' feature to perform this next task (see the corresponding topic video for instructions).

If you are using 'Excel for the web' (the online version of Excel), the 'Text to Columns' feature is not available, but you can achieve the same results using functions, as shown in the steps below.

### Task B: Use LEFT, RIGHT, LEN, and SEARCH functions to clean data:

1. Select column **A (Cust\_Name)**, right-click and choose **Insert Columns**.
2. Select column **A** again, right-click and choose **Insert Columns**.
3. In cell **A1**, type **Customer\_Firstname** and in cell **B1**, type **Customer\_Lastname**.

4. Click **C1**, then on the **Home** tab, click **Format Painter**, then drag across to **A1** and **B1**.
5. Double-click the **divider between columns A and B**.
6. In cell **A2** type `=LEFT(C2, SEARCH( " ",C2,1))` and press **Enter**.
7. In cell **B2** type `=RIGHT(C2,LEN(C2)-SEARCH( " ",C2,1))` and press **Enter**.
8. Double-click the **Fill Handle** on cell **A2**.
9. Double-click the **Fill Handle** on cell **B2**.

**Congratulations! You have completed Lab 5, and you are ready for the next topic.**

# Author(s)

---

- Sandip Saha Joy

# Other Contributor(s)

---

- Steve Ryan

# Changelog

---

Date	Version	Changed by	Change Description
2020-09-10	1.2	Steve Ryan	Added software/dataset info
2020-07-07	1.1	Steve Ryan	ID/Tech review pass
2020-07-01	1.0	Sandip Saha Joy	Initial version created

---

**© IBM Corporation 2020. All rights reserved.**