Chapter 2

# Applying the Concepts: Excel Instructions

## Identify Missing Data

1. Open the Titantic-Dataset.csv dataset in Excel.
2. Verify that the PassengerId variable is in Column A and Age is in Column F.
3. Identify the total number of passengers:
   1. Use the Count function to count the number of values.
      1. In cell O3, type:

=COUNTA(A:A)-1

* + 1. Note: we subtracted 1 to exclude the column's title (row 1).

1. Identify missing age values:
   1. Use the Count Blank function to count the number of missing values.
      1. In cell O4, type:

=COUNTBLANK(F1:F892)

1. Calculate the percentage of missing data:
   1. Use a formula to calculate the percentage of missing age values in an empty cell.
      1. In cell O5, type:

=(O4/O3)\*100

* 1. Divide the number of missing age values by the total number of passengers.
  2. Convert to a percentage.

## Impute Missing Values From Internal Data

1. Open the World\_Happiness\_Report\_2005-2021.csv data set in Excel.
2. Verify the variable Healthy life expectancy at birth is in Column F.
3. Calculate mean and median before imputation:
   1. In cell O1, type:

=AVERAGE(F:F)

* 1. In cell O2, type:

=MEDIAN(F:F)

1. Perform mean imputation:
   1. Insert a new column next to the Healthy life expectancy at birth column for the imputed values. You can do this by highlighting Column G and Right-Click → Insert.
   2. Title this new column “Mean-imputed healthy life expectancy at birth” in cell G1.
   3. In cell G2, type:

=IF(ISBLANK(F2), O$1, F2)

* 1. Drag this formula down the entire column to apply it to all rows.

1. Calculate via the mean imputation method:
   1. Use the AVERAGE function on the new column to find the mean HLEB after the mean imputation. In an empty cell, type:

=AVERAGE(G:G)

1. Perform median imputation:
   1. Insert a new column next to the mean-imputed HLEB column for the imputed values. You can do this by highlighting Column H and Right-Click → Insert.
   2. Title this new column “Median-imputed healthy life expectancy at birth” in cell H1.
   3. Verify that your median from step 3b above is now in Q2. (NOTE: Your mean from step 3a above also shifted to column Q. Be sure to adjust your formula in G2 accordingly and then drag to the rest of the columns.)
   4. Use an IF statement combined with ISBLANK. In cell H2, type:

=IF(ISBLANK(F2), Q$2, F2)

* 1. Drag this formula down the entire column to apply it to all rows.

1. Calculate via the median imputation method:
   1. Use the AVERAGE function on the new column to find the mean HLEB after median imputation. In an empty cell, type:

=AVERAGE(H:H)

## Identify Extreme Data Values

1. Open birdwatcher.csv as an Excel file.
2. Verify the age variable is in Column B.
3. Calculate the 95th percentile age range.
   1. Highlight the age column and select Sort & Filter → Filter.
   2. Click the gray arrow in the filter dropdown Choose one, and select Top 10.
   3. Change “10” to “5” and change “Item” to “Percent”. Click Ok.
   4. Highlight the age column again and select Sort & Filter → Sort Largest to Smallest. When prompted with a warning, click Continue with the current selection only to sort the filtered data.
   5. Record the largest and smallest age values in this filtered sample of the 95th percentile (top 5%). Write this down, this is your 95th percentile age range.
4. Calculate the 5th percentile age range.
   1. Highlight the age column and select Sort & Filter → Filter.
   2. Click the gray arrow in the filter dropdown Choose one, and select Bottom 10.
   3. Change “10” to “5” and change “Item” to “Percent”. Check Auto Apply or click Apply Filter. Click Ok.
   4. Highlight the age column again and select Sort & Filter → Sort Smallest to Largest. When prompted with a warning, click Continue with the current selection only to sort the filtered data.
   5. Record the smallest and largest age values in this filtered sample of the 5th percentile (bottom 5%). Write this down, this is your 5th percentile age range.
5. Count respondents above age 100 and below age 18:
   1. Insert into an empty cell:

=COUNTIF(B:B,">100")

* 1. Insert into an empty cell:

=COUNTIF(B:B,"<18")

* 1. Record these counts.

1. Save your Excel file after completing these steps.

## Clean Incorrect Formats

1. Open the birdwatcher.csv data set in Excel.
2. Verify that the Favorite\_Bird variable is in Column C.
3. Calculate totals before cleaning the data.
   1. In an empty cell, type:

=COUNTIF(C:C,"sparrow")

* 1. In an empty cell, type:

=COUNTIF(C:C,"woodpecker")

* 1. In an empty cell, type:

=COUNTIF(C:C,"hummingbird")

1. Use FIND AND REPLACE to clean the data.
   1. From the home tab, choose Find & Select → Replace.
   2. In the ‘Find what’ field, type “sparrows”; in the ‘Replace with’ field, type “sparrow”.
   3. Repeat this process, but in the ‘Find what’ field, type “woodpeckers”, and in the ‘Replace with’ field, type “woodpecker”.
   4. Repeat this process for a final time but in the ‘Find what’ field, type “hummingbirds” and in the ‘Replace with’ field, type “hummingbird”.
2. Use an IF statement for further cleaning.
   1. Right-click on Column D and choose Insert.
   2. Title the new blank Column D “Favorite\_Bird\_cleaned”.
   3. Type into an empty cell the following formula to replace words that start with S (s-) and end with W (-w) to “sparrow”, words that start with W (w-) and end with R (-r) to “woodpecker”, and words that start with H (h-) and end with D (-d) to “hummingbird”:

=IF(AND(LEFT(C2,1)="s",RIGHT(C2,1)="w"),"sparrow",IF(AND(LEFT(C2,1)="w",RIGHT(C2,1)="r"),"woodpecker",IF(AND(LEFT(C2,1)="h",RIGHT(C2,1)="d"),"hummingbird",C2)))

* 1. Drag this formula down to apply it to all rows.

1. Calculate totals after cleaning the data.
   1. In an empty cell, type:

=COUNTIF(D:D,"sparrow")

* 1. In an empty cell, type:

=COUNTIF(D:D,"woodpecker")

* 1. In an empty cell, type:

=COUNTIF(D:D,"hummingbird")

## Create a Categorical Variable From a Quantitative Variable

1. Open the AmesHousing.csv file in Excel.
2. Verify that the Sales Price variable is in Column CD
3. Add a new column to create the Sales Price categories:
   1. Right-click on Column CE and choose Insert.
   2. Title the new blank Column CE “SalePrice\_Cat”.
4. Use an IF statement to categorize Sales Price values:
   1. In Cell CE2 type:

=IF(CD2<=100000,"Cheap",IF(AND(CD2 >=100001,CD2<= 200000),"Moderate","Expensive"))

* 1. Drag this formula down to the end of the data and apply it to all rows.

1. Count the number of homes in each category using the COUNTIF function:
   1. In an empty cell, type:

=COUNTIF(CE:CE, "Expensive")

* 1. In an empty cell, type:

=COUNTIF(CE:CE, "Moderate")

* 1. In an empty cell, type:

=COUNTIF(CE:CE, "Cheap")

1. Add a new column to create different Sales Price categories:
   1. Right-click on Column CF and choose Insert.
   2. Title the new blank Column CF “SalePrice\_Cat2”.
2. Use an IF statement to modify the categories and split the “moderate” category:
   1. In Cell CF2, type:

=IF(CD2<=100000,"Cheap",IF(AND(CD2 >=100001,CD2<=150000),"Cheap Moderate",IF(AND(CD2>150000,CD2<=200000),"Expensive Moderate","Expensive")))

* 1. Drag this formula down to the end of the data and apply it to all rows.

1. Count the number of homes in each newly modified category using the COUNTIF function:
   1. In an empty cell, type:

=COUNTIF(CF:CF, "Expensive")

* 1. In an empty cell, type:

=COUNTIF(CF:CF, "Expensive Moderate")

* 1. In an empty cell, type:

=COUNTIF(CF:CF, "Cheap Moderate")

* 1. In an empty cell, type:

=COUNTIF(CF:CF, "Cheap")

## Create a Boolean Variable that Represents a Categorical Variable

1. Open the zoo.csv file in Excel.
2. Verify the class\_type variable is in Column R.
3. Add a new column to create bird, mammal, and reptile variables:
   1. Right-click on Column S and choose Insert. In Cell S1, type: “bird”.
   2. Right-click on Column T and choose Insert. In Cell T1, type: “mammal”.
   3. Right-click on Column U and choose Insert. In Cell U1, type: “reptile”.
4. Use IF statements to determine whether each animal falls into that column’s category:
   1. Type the following formulas into empty cells:
      1. In cell S2, type:

=IF(R2=2, TRUE, FALSE)

* + - 1. Drag this formula down to the last row of the data set.
    1. In cell T2, type:

=IF(R2=1, TRUE, FALSE)

* + - 1. Drag this formula down to the last row of the data set.
    1. In cell U2, type:

=IF(R2=3, TRUE, FALSE)

* + - 1. Drag this formula down to the last row of the data set.
  1. Note, in the class\_type column 2 represents birds, 1 represents mammals, and 3 represents reptiles.
  2. Copy these formulas down to the last row of the data set.

1. Type the following formulas into empty cells:
   1. Bird:

=COUNTIF(S:S, TRUE)

* 1. Mammal:

=COUNTIF(T:T, TRUE)

* 1. Reptile:

=COUNTIF(U:U, TRUE)

## Splitting a Variable Into Multiple Variables

1. Open the birdwatcher.csv file in Excel.
2. Verify the Date variable is in Column F.
3. Add a new column to create the month\_of\_entry variable:
   1. Right-click on Column G and choose Insert.
   2. In Cell G1, type: “month\_of\_entry”.
4. Extract the month of entry from the date variable using the TEXT function:
   1. In Cell G2, type:

=TEXT(F2,"mmmm")

* 1. Copy the formula down.

1. Use the COUNTIF function to determine the distribution of monthly entries:
   1. Type into an empty cell:

=COUNTIF(G:G, "January")

1. Repeat this process for the remaining 11 months using the same formula but replacing the month's name.