TT

TEXTBOOK TESTBANKS

Solutions Manual for Data Science For All 1st Davis

SOLUTIONS MANUAL SAMPLE PREVIEW

PUBLISHER

Pearson

COPYRIGHT

2026

RESOURCE TYPE

Solutions Manual

ISBN

9780135311400

Part II

1. Data Collection
2. Data Preparation
3. Data Analysis
4. Descriptive
5. Diagnostic
6. Storytelling
7. Predictive
8. Prescriptive
9. Data science lifecycle

Chapter 2: Data Wrangling: Preprocessing

Chapter Review Questions

Section 1: What Is Data Wrangling?

1. What term refers to the process of converting raw data into a tidy, organized format suitable for analysis?
   Data wrangling

2. What are the three key components of data wrangling?
   Cleaning, transforming, integrating

3. Explain the data wrangling process.
   Answers may vary. Example: In data wrangling, the data scientist takes collected data and prepares it for analysis or storytelling, or both. It involves 3 stages: cleaning the data, transforming the data, and integrating the data.

4. What is the first step in data wrangling?
   Cleaning the data

5. In the context of data wrangling, what does “integrating the data” refer to?
   Combining multiple data sets

6. What is the purpose of cleaning data in data wrangling?
   To enhance the quality of the data for accurate analysis

7. What does the term “dataframe” mean in the context of data wrangling?
   A tabular data structure with named columns

Section 2: Cleaning Missing Data

8. What are three ways of dealing with missing data values?
   Removing observations with missing values; imputation from internal data; imputation from external data.

9. What is an advantage of imputing missing values over leaving them missing?
   Bigger sample size, because you do not need to remove missing values for analyses

10. What issue might arise when dropping rows with any missing values from a dataset?
    Answers may vary. Example: Biased sample

11. When cleaning data, under which condition would it make most sense to remove an entire row of data?

Duplicate observations

12. What term refers to estimating missing values based on other available information?
    Imputing

13. Describe a potential issue that arises from imputing missing data values.
    Answers may vary. Example: Influencing analyses with artificial data

14. What is the mean GPA, ignoring only the missing GPA value?
    3.22

15. What is the mean GPA, ignoring any row with a missing value?
3.25

16. Impute the CA student's missing age by taking the mean age of other CA students. What is their imputed age?
    20 (mean of other CA students)

17. Impute the missing favorite color by the most frequent favorite color of students with the same GPA. What is the imputed color?
    Blue (mode for students with 3.3 GPA)

Section 3: Cleaning Anomalous Values

18. What term refers to data values that are conspicuously different from other values of the same variable?
    Outliers

19. If a data set contains a person's age as 2995, how might this be addressed in data cleaning?
    Answers may vary. Sample answer: Average all ages and replace it with the mean

20. In data cleaning, what term refers to values that are clearly invalid or impossible?
    Implausible values

21. Describe an example of incorrect formatting in data.
    Answers may vary. Example: Age entered as text

22. What is the best way to handle an extreme data value that is verified to be correct?
    Answers may vary. Example: Keep it

Section 4: Transforming Quantitative Variables

23. What term refers to creating a new categorical variable by assigning categories to intervals of an existing quantitative variable?
    Categorization

24. How might a quantitative variable, such as annual income measured in dollars, be transformed for easier visualization?
    Convert to a scale such as income measured in thousands of dollars

25. Describe an example of creating a new quantitative variable by combining two existing quantitative variables.
    Answers may vary. Example: Calculating BMI from weight and height

26. How does scaling a quantitative variable by dividing its values help create better visualizations?
    It reduces the number of digits.

27. In general, what is the best practice when transforming variables?
To create new variables leaving originals intact

28. What can be done to make a quantitative variable with a wide range easier to visualize and interpret without distorting the data?
Answers may vary. Example: Rescale the variable by a constant.

29. How does creating a quantitative variable from a categorical one enable more types of analysis?
It allows for different numerical summary statistics.

30.

student	**data_usage (MB)**	**major**	**data_usage (GB)**
Jackson	10587	Art	10.34
Oneal	15789	CS	15.42
Ortega	12678	Bio	12.38
White	14567	Math	14.23
Martinez	16987	Eng	16.59
Dixon	12356	Psych	12.07
Garcia	13489	Hist	13.17
Harris	15782	Econ	15.41
Brown	14256	Socio	13.92
Baker	13567	Music	13.25

Mean data usage in Gigabytes: 13.68 GB
Median data usage in Gigabytes: 13.59 GB

31.

student	**height (in)**	**major**	**height (cm)**

Washington	71	Football	180.34
Jefferson	74	Basketball	187.96
Patel	68	Soccer	172.72
Gupta	69	Baseball	175.26
Li	74	Tennis	187.96
Kim	75	Swimming	190.50
Rodriguez	73	Track	185.42
Garcia	72	Hockey	182.88
Jackson	66	Wrestling	167.64
Sharma	70	Golf	177.80

Mean height in centimeters: 180.85 cm

32. In data cleaning, what might be done with text values in a quantitative variable like height?
Convert to numeric equivalent

33. Create a new variable, BMI, by filling in the values of the final column's cells using the formula (rounded to one decimal place):
bmi = weight (lb) / [height (in)]^2 x 703

student	**weight (lb)**	**height (in)**	**major**	**bmi**
A	120	62	Math	21.9
B	135	64	Bio	23.2
C	150	66	CS	24.2

D	165	68	Eng	25.1
E	180	70	Psych	25.8
F	155	64	Hist	26.6
G	160	66	Art	25.8
H	175	68	Econ	26.6
I	190	70	Socio	27.3
J	205	72	Music	27.8

34. What is the mean BMI for this data set (rounded to one decimal place)?
Mean BMI = 25.4

35. Calculate the cost per mile for each trip using the formula:
Cost per Mile ($) = (Fuel Consumption (gallons) x Cost per Gallon ($)) / Distance (miles)

student_id	**distance (miles)**	**fuel_consumption (gallons)**	**cost_per_gallon ($)**	**major**	**cost_per_mile ($)**
345634	50	2	3	Art	0.12
534534	100	4	2.5	CS	0.10
832344	150	5	3.5	Bio	0.12
242345	200	6	3	Math	0.09
859285	250	7	2.5	Eng	0.07
894993	100	3	3	Psych	0.09

388539	150	4	2.5	Hist	0.07
393559	200	5	3.5	Econ	0.09
395222	250	6	3	Socio	0.07
825234	300	7	2.5	Music	0.06

36. What is the mean cost per mile? Use two decimal places for all answers.
    Mean cost per mile: $ 0.09

37. Calculate the force (in newtons) applied to each object of the final column using the formula (to one decimal place):
    force = mass x acceleration

object	**mass (kg)**	**acceleration (m/s^2)**	**force (N)**
1	5	0.5	2.5
2	10	1	10.0
3	15	1.5	22.5
4	20	2	40.0
5	25	2.5	62.5
6	30	3	90.0
7	35	3.5	122.5
8	40	4	160.0
9	45	4.5	202.5

10	50	5	250.0

38. What is the mean force applied to the objects (one decimal place)?
Mean Force: 96.3 N

39. Create a categorical age_group variable using the following categories:
18-35: Young Adult
36-55: Middle Aged
56+: Senior

participant	age	height	weight	age_group
1	32	5'11	175	Young Adult
2	55	5'5	132	Middle Aged
3	62	5'9	195	Senior
4	18	5'3	120	Young Adult
5	23	6'1	180	Young Adult
6	43	5'7	150	Middle Aged
7	51	5'8	170	Middle Aged
8	28	5'4	135	Young Adult
9	47	6'0	200	Middle Aged
10	61	5'2	142	Senior

40. How many participants would be in the Young Adult category?
4

41. How many would be Middle Aged?
4

42. How many would be Senior?
2

Section 5: Transforming Categorical Variables

43. How might the uncommon (low frequency) categories of a categorical variable be simplified, given there are no concerns about othering?
Merging uncommon categories into one new, slightly larger category

44. What term refers to extracting multiple variables from a single categorical variable?
Splitting

45. Why is it important to assign every possible value to a category when creating categories from a quantitative variable?
It ensures that every value is properly assigned and prevents missing values.

46. What is a Boolean variable? When is it used?
It is a variable with two values, usually true or false. It is used to encode categorical data, converting it to a binary form.

47. What are some other names for Boolean variables?
Indicator variables, binary variables, dummy variables, logical variables, dichotomist variables.

48. When encoding a categorical variable numerically, why is the choice of values important?
The differences between the quantitative values should mirror differences between categories.

49. What term refers to reducing the number of distinct values a categorical variable can take?
Condensing

50. Describe the process of condensing a categorical variable.
Answers may vary. Example: the process involves merging small categories into broader ones, which can simplify the variable for analysis while reducing its informational detail.

51. When would splitting a categorical variable into multiple new variables be useful?
When a single value encompasses multiple distinct attributes

52. Provide three examples of categorical variables.
Answers may vary. Examples: college major, type of donation; animal species; country

53. What term is used to refer to the creation of quantitative variables from a categorical variable?
Encoding

54. What is an example of creating a categorical variable from a quantitative one?
Answers may vary. Example: Converting test scores to letter grades.

55. What is a potential mistake that can arise when encoding categorical data numerically?
Answers may vary. Example: Numerical codes not aligning with the meaning of categories.

56. How many people have January as their birth month?
1 (January)

57. How many were born in the last 15 days of a month?
5 (Last 15 days)

58. What is the mean birth year, rounded to the nearest year?
1989 (Mean year)

59. Recode the responses as follows:
Strongly Disagree = 1 | Disagree = 2 | Neutral = 3 | Agree = 4 | Strongly Agree = 5

participant	**response**	**age**
1	1	35
2	4	29
3	3	42

4	5	27
5	4	38
6	2	31
7	3	44
8	4	24
9	5	36
10	1	39

60. What is the mean of the recorded responses rounded to one decimal place?
Mean: 3.3

61. What is the median?
Median: 3.5

62. Split the major into Department and Major columns:

student	**major**	**department**	**major**
1	Computer Science - Software Engineering	Computer Science	Software Engineering
2	Physics - Applied Physics	Physics	Applied Physics

3	Business - Accounting	Business	Accounting
4	History - American History	History	American History
5	Engineering - Electrical Engineering	Engineering	Electrical Engineering
6	English - Creative Writing	English	Creative Writing
7	Chemistry - Biochemistry	Chemistry	Biochemistry
8	Education - Elementary Education	Education	Elementary Education
9	Math - Applied Math	Math	Applied Math
10	Nursing - Nursing	Nursing	Nursing

63. How many students have Business majors?
1

Section 6: Reshaping a Dataset

64. Describe the process of converting a data set from wide format to long format.
Answers may vary. Example: It involves increasing the number of rows and reducing columns in the data set.

65. Why might a data scientist prefer a long-format dataset?
It is better for computing systems to perform analysis.

Section 7: Combining Datasets

66. What is an advantage of stacking two datasets with the same variables but different observations?
It increases the number of observations.

67. Which data integration technique combines datasets with all the same variables but additional observations?

Stacking

68. When would it be appropriate to augment two datasets by binding them horizontally?
Answers may vary. Example: When the data sets contain the same observations but different variables

69. What does “merging datasets” typically involve?
Answers may vary. Example: When two datasets are merged, they will have at least one common variable. Observations between the two datasets are matched on at least one of their common variables.

70. For which type of data integration must datasets have at least one common variable?
Merging

71. When merging two datasets, what must they have in common?
At least one variable

72. What term refers to binding two datasets together horizontally when they have the same observations but completely different variables?
Augmenting

73. What is the primary purpose of combining datasets?
To facilitate more comprehensive analysis by integrating diverse data

74. If these tables were augmented by binding them horizontally, how many rows and how many columns would the resulting table have? (Don’t count the column headings as rows.)
3 rows, and 6 columns.

75. If the graduate student stacks these data tables, how many rows would the new stacked table contain?
6 rows: one each for Amanda, Diego, Sara, Tia, Anna, and Sarah (which is distinct from Sara)

76. How many columns would the new stacked table have?
4 columns: name, math_test_1, math_test_2, section

77. If the professor stacks these two data tables, how many rows would the stacked table have?
4 rows: one each for Diego’s Semester 1, Sam’s Semester 2, Alex’s Semester 1, and Alex’s Semester 2.

78. How many columns would the stacked table have?
5 columns: name, exam_1, exam_2, final_exam, semester

79. If the teacher merges these tables using Name, how many rows would the new table have?
4 rows: one each for Anna, Diego, Sara, and Tia

80. How many columns would it have?
6 columns: name, reading_test_1, reading_test_2, reading_test_3, math_test_1, math_test_2

81. If they merge the datasets, how many unique Customer ID values would exist in the merged dataset?
3 Customer IDs

82. How many rows would the merged dataset contain?
3 rows (corresponding to the 3 Customer IDs)

83. Merge the categories of the categorical variable, major, into 3 categories:
- STEM: Biology, Chemistry, Physics, Math, Computer Science, Engineering
- Humanities: English, History, Arts
- Other: Any other majors

student	major	gpa	clubs

1	STEM	3.2	1
2	Humanities	3.0	2
3	STEM	3.5	0
4	Humanities	3.2	1
5	STEM	3.8	0
6	STEM	3.4	2
7	Humanities	3.1	3
8	STEM	3.7	1
9	Humanities	3.0	0
10	STEM	3.5	2
11	STEM	3.6	1
12	STEM	3.2	0
13	Humanities	3.9	2
14	STEM	3.3	1
15	Humanities	2.9	0
16	STEM	3.8	2
17	Humanities	3.0	3
18	STEM	3.5	1
19	STEM	3.7	0
20	Humanities	3.4	2

84. How many students have a STEM major?
12 (STEM)

85. How many have a Humanities major?
8 (Humanities)

86. How many have an Other major?
0 (Other)

87. Merge the categories of the categorical variable, Crop, into 2 categories:
- Row crops: Corn, Soybeans, Cotton
- Small grains: Wheat

farm	**crop**	**acres**	**yield**
1	Row crops	80	8500
2	Row crops	160	9200
3	Small grains	240	11,000
4	Row crops	100	9000
5	Row crops	200	10,000
6	Row crops	150	8000
7	Small grains	300	12,000
8	Row crops	90	8000
9	Row crops	140	7500
10	Row crops	180	9500
11	Small grains	220	10,500
12	Row crops	160	8500
13	Row crops	110	9500
14	Row crops	170	9000
15	Small grains	260	11,500

88. How many farms grow row crops?
11 (row crops)

89. How many farms grow small grains?
    4 (small grains)

Explaining the Concepts

1. Pieces of information provided on some of these forms might include name, address, age, gender, or ID number.

   Different people might fill out their names differently. For example, one person might put their first name followed by their last name. Another person might put their last name first followed by their first name. Similarly, different people might fill out their age differently. One person might use integers while another might write out the number, *20* vs *twenty*.

   Changing prompts on a form can help make the data more consistent. Some changes can include providing examples or detailed instructions.

   These issues should be handled at each stage of the data pipeline, including collection and cleaning.

2. For categorical data, standardize the labels/names of the categories. For example, if there is no difference in the meaning between *Dog, dog,* and *DOG*, then they should be standardized when cleaning the data. Lowercasing (*dog)* or title casing (*Dog)* are commonly used standards. For quantitative data with decimal points, e.g. *10.5* or *3.45*, keep the number of digits consistent. For example, we might want to replace *10.5* with *10.50.*

3. Retaining variables is a good practice that helps ensure the work is reproducible. Additionally, since data science is an iterative process, replacing variables might remove some information that we will only realize later is important. We might not be able to recover the original data if we replace the original variables.

   When replacing variables by condensing variables, we might remove information. For example, merging *pet_preference* into just “dog”, “cat”, “none”, or “other” and removing *pet_preference* would prevent us from keeping track of the other responses (Section 2.5). Categorization is another example of a data transformation method that can result in a loss of information if we replace a variable. For example, replacing the ages with “child”, “adult”, or “senior” and removing the original age will prevent us from computing quantitative statistics, e.g. the mean or median age.

4. *Cleaning* and *transforming* the data of each dataset separately first can help integrate data correctly by making the datasets match each other better. For example, imagine that the seasons in Table 2.10 are in lowercase, e.g. *“fall”*. When augmenting Tables 2.10 and 2.11 together, we would not be able to identify that *average_rainfall* and *average_temperature* are variables for the same observation.

5. Detailed instructions and explanations of Mirabel’s data-wrangling steps would help Miles reproduce Mirabel's work. Also, if Miles wants to compare a new group of students with the students in Mirabel’s group, Miles will need to repeat the exact same process that Mirabel used. This process would include the exact data wrangling steps.

   Ensuring that work can be replicated on her data makes Mirabel confident in the results of her analysis. Ensuring that the work can be applied accurately to new data enables researchers to continue Mirabel’s work or use the same approach to study another population.

Try It Yourself

Try It Yourself: Identifying Missing Data (2.2)

1. The number of missing values in the GPA column is 596.

2. The total number of rows in the dataset is 5187.
3. The percentage of rows with missing GPA values is approximately 11.49%, a high amount that should elicit some concern.
4. STAR Framework:
    - Scope the Data's Context: The dataset under consideration includes GPA data alongside other variables, such as voting behavior for students. A substantial number of missing GPA values were identified, which is critical for understanding the educational background and its possible association with voting behavior.
    - Track the Question behind the Exploration: The primary objective is to assess the extent of missing GPA records to evaluate the dataset's reliability, and to inform potential data cleaning or imputation strategies. This assessment is crucial to ensure the accuracy and relevance of any subsequent analysis involving GPA data.
    - Articulate the Results: In the dataset comprising 5187 rows, there are 596 missing values in the GPA column. This constitutes approximately 11.49% of the dataset, a noticeable proportion that could impact the validity of any analysis that relies on GPA data.
    - Respond with Next Steps: Given the high percentage of missing GPA data, it is essential to consider methods for handling these gaps, such as data imputation, or to adjust the analysis approach to account for this limitation. Further exploration into the reasons behind these missing values might also provide insights into data collection processes and help improve data quality for future studies.

Try It Yourself: Impute Missing Values from Internal Data (2.2)

1. The mean GPA is approximately 2.98.
2. The median GPA is 3.00.
3. The mean GPA remained approximately 2.98.
4. The median GPA changed to approximately 2.98.
5. STAR Framework:
    - Scope the Data's Context: The voter survey dataset, which includes GPA data, has missing values in the GPA column. The mean imputation method is used to address these missing values, aiming to maintain the integrity of the dataset for reliable analysis.
    - Track the Question Behind the Exploration: The exploration focuses on understanding the impact of replacing missing GPA values with the mean GPA. The key questions involve assessing the mean and median GPA before and after the imputation process to gauge how this method affects these central measures.
    - Articulate the Results: Initially, the mean GPA in the dataset was approximately 2.98, and the median GPA was 3.0. After employing mean imputation to replace missing values, the mean GPA remained around 2.98. However, the median GPA shifted to approximately 2.98, showing a slight change due to the imputation process.
    - Respond with Next Steps: The results suggest that while mean imputation preserves the mean GPA, it slightly alters the median GPA. This finding should be considered when analyzing the dataset post-imputation, particularly in studies where the median GPA is a critical measure. Further, exploring other imputation methods or considering the implications of missing data on the overall analysis would be beneficial for comprehensive understanding and interpretation.

Try It Yourself: Identify Extreme Data Values (2.3)

1. The oldest 5% of respondents are between 22 and 115 years old.
2. The youngest 5% of respondents are between 3 and 18 years old.
3. There are three respondents who are above 100 years old.
4. There are nine respondents who are below 18 years old.
5. STAR Framework:
    - Scope the Data's Context: In evaluating the vote dataset, primarily consisting of university campus respondents, our aim is to scrutinize the age data for potential inaccuracies. Identifying extreme age

values is crucial for maintaining data quality and ensuring reliable analysis, especially considering university policies regarding minor participation.

- Track the Question Behind the Exploration: Our exploration is centered on determining the ages of the oldest and youngest 5% of respondents, identifying individuals above 100 years old, and those below 18. This analysis helps in discerning whether these age values are plausible or constitute outliers.
- Articulate the Results: The analysis reveals that the ages of the oldest 5% of respondents range from 22 to 115 years, and the youngest 5% range from 3 to 18 years. There are three respondents over 100 years old and nine below 18. The presence of respondents below 18 is an outlier, as the university prohibits their participation as minors. While respondents above 100 are unlikely on a university campus, it's less clear-cut and may warrant a more inclusive approach.
- Respond with Next Steps: Given these findings, the immediate step involves addressing the presence of minors in the dataset, as they represent clear data entry errors or policy violations. For the respondents above 100, a deeper investigation might be necessary to verify the authenticity of these entries. Adjusting the dataset by removing or further scrutinizing these extreme values will enhance its validity and reliability for subsequent analyses.

Try It Yourself: Clean Incorrect Formats (2.3)

1. Before the basic data cleaning, in the dataset:
   a. The number of people who indicated their favorite pet as "dog" was 1646.
   b. The number of people who indicated their favorite pet as "cat" was 1808.
2. After the basic data cleaning, the numbers changed as follows:
   a. The number of people who indicated their favorite pet as "dog" increased to 1833.
   b. The number of people who indicated their favorite pet as "cat" increased to 2004.
2. STAR Framework:
   - Scope the Data's Context: The focus is on the Favorite_pet[AS1] column in the vote dataset. The goal is to standardize the data by correcting variations in the entries for favorite pets, specifically targeting the categories "dog" and "cat."
   - Track the Question Behind the Exploration: The exploration aims to quantify the impact of data cleaning on the representation of "dog" and "cat" in the dataset. This involves counting the occurrences of these pets before and after applying specific cleaning rules to unify similar entries under standard labels.
   - Articulate the Results: Before cleaning, the count for "dog" was 1646 and for "cat" was 1808. After applying cleaning rules—standardizing variations of "dogs" to "dog," "cats" to "cat," and including words starting with "d" and ending in "g" under "dog," and those starting with "c" and ending in "t" under "cat" — the count for "dog" increased to 1833 and for "cat" to 2004.
   - Respond with Next Steps: The results indicate that the cleaning process successfully identified and corrected inconsistent entries, leading to a more accurate representation of pet preferences in the dataset. Establishing guidelines or using tools for data entry can prevent such inconsistencies in future data collection.

Try It Yourself: Create a Categorical Variable from a Quantitative Variable (2.4)

1. There are 2349 students with a high GPA (with values from 3.00 to 4.00), and 2242 students with a low GPA (with values from 2.00 to 2.99) in the dataset.
2. STAR Framework:
   - Scope the Data's Context: The analysis involves the GPA variable in the vote dataset, with an objective to categorize GPAs into two groups for a simplified analysis. The new variable, GPA_category, is created to classify GPAs as "low" for the values of 2.00–2.99 and "high" for values of 3.00–4.00.
   - Track the Question Behind the Exploration: The exploration seeks to understand the distribution of GPAs among students in terms of these two categories. This categorization helps in analyzing the

data more effectively by reducing the complexity of continuous GPA values into simpler, categorical terms.

- Articulate the Results: The categorization resulted in identifying 2349 students with a "high" GPA (with values from 3.00 to 4.00), and 2242 students with a "low" GPA (with values from 2.00 to 2.99). This distinction provides a clearer picture of the overall academic performance distribution within the dataset.
- Respond with Next Steps: Given these results, further analysis could explore associations between GPA categories and other variables in the dataset, such as voting behaviors or other demographic factors. Additionally, considering different GPA categories could offer alternative perspectives on the data.

Try It Yourself: Create a Boolean Variable that Represents a Categorical Variable (2.5)

1.
    a. 2081 drinking_age values are 1.
    b. 3106 drinking_age values are 0.
2.
    a. Number of people of drinking age who voted: 1199
    b. Number of people not of drinking age who voted: 1715
    c. Percentage of drinking age people who voted: 57.6%
    d. Percentage of non-drinking age people who voted: 55.2%
3. STAR Framework:
    - Scope the Data's Context: A drinking_age Boolean variable was added to the dataset to differentiate between individuals above and below the legal drinking age, facilitating an analysis of voting behavior across different age groups.
    - Track the Question Behind the Exploration: The analysis focused on determining how many individuals fell into each drinking age category and assessed their respective voting participation rates.
    - Articulate the Results: In the dataset, 2081 individuals were classified as of legal drinking age (21 and over), with 1199 having voted, representing 57.6% of this group. In contrast, 3106 individuals were under the drinking age, with 1715 having voted, making up 55.2% of this younger group.
    - Respond with Next Steps: These findings suggest a marginal difference in voting rates between the two age groups. Further investigation could involve exploring demographic and psychographic factors that might influence these voting behaviors, providing deeper insights for voter engagement strategies.

Try It Yourself: Splitting a Variable into Multiple Variables (2.5)

1.
    - January: 474 people
    - February: 398 people
    - March: 426 people
    - April: 432 people
    - May: 436 people
    - June: 406 people
    - July: 432 people
    - August: 437 people
    - September: 443 people
    - October: 433 people
    - November: 459 people
    - December: 411 people

2. STAR Framework:

- Scope the Data's Context: In the voter survey dataset, we isolate the birth month from the existing birthdate information, creating a new column for this purpose. This step enables an analysis focused on discerning potential associations between voters' birth months and their voting behaviors.
- Track the Question Behind the Exploration: The primary question centers on quantifying the distribution of voters across different birth months to explore if and how the birth month might be associated with voting patterns.
- Articulate the Results: The breakdown by month revealed varying numbers of people born in each month, with January having the highest at 474 and February the lowest at 398. The distribution is relatively even across the year, with most months having between 400 and 450 people.
- Respond with Next Steps: Given this distribution, the next phase of analysis can delve into comparing these birth month groupings with voting behaviors. This could uncover trends or patterns that are specific to certain times of the year, potentially offering insights into voter behavior associated with demographic factors such as age.

Applying the Concepts

Activity: Identify Missing Data

1. Overall, there were 891 passengers recorded in this dataset.
2. 177 passengers were missing a value for the age variable.
3. As a percentage, this means that rounded to the nearest whole percentage point, 20% of all passengers did not have their age recorded.
4. If child passengers were less likely to have their age recorded in this dataset than adult passengers, then the average age we calculate would be higher than the true average age.
5. This strategy for handling missing data is called imputation, where the missing values are filled with an educated guess. This guess could be made using means, medians, or prediction models. Deciding on the most appropriate method requires careful consideration.

Activity: Impute Missing Values from Internal Data

1. The mean imputation method produces an average life expectancy of 63.18 years.
2. The median imputation method produces an average life expectancy of 63.23 years.

Activity: Identify Extreme Data Values

1. The youngest age of the oldest 5% is 101
2. The oldest age for the youngest 5% is 22
3. There are 1,333 birdwatchers who are above 100 years old.
4. There are 11 birdwatchers who are below 18 years old.
5. There are at least four impossible age values: -1, 0, 134, 723. Validity of values from 3 to 9 can also be questioned.
6. STAR Framework:
    - Scope the Data's Context: The birdwatcher dataset contains information about birdwatchers, including their ages. Our focus is on understanding the distribution of ages, especially identifying any extreme or impossible values.
    - Track the Question behind the Exploration: We aim to uncover if there are any unusually high or low ages, as these could be errors affecting data integrity. Specifically, we want to know the percentile values of the oldest and youngest 5% of birdwatchers, the number of birdwatchers above 100 and below 18, and any impossible age values. Identifying and addressing extreme data

values in the birdwatcher dataset ensures data accuracy and reliability, facilitating informed decision-making and robust statistical analyses.

- Articulate the Results: The analysis reveals that the youngest age of the oldest 5% of birdwatchers is 101 years, and the oldest age of the youngest 5% is 22 years. There are 1,333 birdwatchers above 100 years and 11 below 18. Additionally, 4 impossible values were found: a negative age, 0, 134, and 723, were found. These results suggest significant data quality issues, particularly with impossibly high and low age values.
- Respond with Next Steps: Further investigation is needed to validate the data source and correct these anomalies. Cleaning the data set by removing or correcting these extreme values is crucial for reliable analyses. Additionally, setting up data validation rules to prevent future data entry errors could be beneficial.

Activity: Clean Incorrect Formats

1. n/a
2.
    a. “sparrow” Answer: 7232
    b. “woodpecker” Answer: 7142
    c. “hummingbird” Answer: 7109
3.
    a. “sparrow” Answer: 8019
    b. “woodpecker” Answer: 7950
    c. “hummingbird” Answer: 7924
4. STAR Framework
    - Scope the Data’s Context: The birdwatcher dataset from an Airbnb birdwatching experience is being cleaned for analysis. The focus is on the Favorite_Bird column, standardizing entries like “sparrows” to “sparrow”, “woodpeckers” to “woodpecker”, and “hummingbirds” to “hummingbird”.
    - Track the Question behind the Exploration: The objective is to determine how basic data cleaning affects the count of participants' favorite birds, specifically “sparrow”, “woodpecker”, and “hummingbird”, to understand better the relationship between bird preferences and interest in repeat experiences.
    - Articulate the Results: Before cleaning, the counts were 7232 for “sparrow”, 7142 for “woodpecker”, and 7109 for “hummingbird”. Post-cleaning, the counts increased to 8019 for “sparrow”, 7950 for “woodpecker”, and 7924 for “hummingbird”. This indicates that cleaning helped rectify inconsistent entries and provided a more accurate representation of bird preferences.
    - Respond with Next Steps: With the data cleaned, further analysis can be conducted to explore any correlations between participants' favorite birds and their willingness to sign up for another tour. Understanding these preferences can aid the guide in tailoring future birdwatching experiences and improving participant satisfaction.

Activity: Create a Categorical Variable from a Quantitative Variable

1. Most houses have prices greater than $100,000 but less than $200,000.
2. n/a
3. The number of houses in the “Cheap Moderate” category is 1019.
4. The largest category is now “Cheap Moderate”, followed by Expensive and Expensive Moderate
5. There are 857 “Expensive” houses.
6. There are 802 “Expensive Moderate” houses.

Activity: Create a Boolean Variable that Represents a Categorical Variable

1. There are 41 mammals in the dataset.
2. There are 20 birds and 5 reptiles in this dataset.

Activity: Splitting a Variable into Multiple Variables

1.
    - January: 2016 entries
    - February: 1965 entries
    - March: 1996 entries
    - April: 1973 entries
    - May: 1947 entries
    - June: 1960 entries
    - July: 2085 entries
    - August: 1972 entries
    - September: 1959 entries
    - October: 1986 entries
    - November: 2006 entries
    - December: 2028 entries
2. STAR Framework:
    - Scope the Data's Context: The birdwatcher dataset was modified to include a new column representing the month, extracted from the existing Date variable. This change aims to investigate seasonal birdwatching patterns and the correlation between the months and types of birds observed.
    - Track the Question behind the Exploration: The exploration sought to quantify the number of birdwatching entries each month, enabling an analysis of how birdwatching experiences and preferences vary seasonally.
    - Articulate the Results: The data revealed a relatively even distribution of birdwatching entries throughout the year, with the highest in July (2,085 entries) and the lowest in May (1,947 entries). This suggests that birdwatching activity is consistently popular across different months, with slight variations.
    - Respond with Next Steps: These findings provide a foundation for further analysis to explore potential correlations between specific months and the types of birds observed. Such insights could be invaluable for understanding the seasonal preferences of various bird species and enhancing the birdwatching experience for participants.

Chapter 3: Making Sense of Data Through Visualization

Chapter Review Questions

Section 1: The Grammar of Graphics

1. What does the term "grammar of graphics" refer to in data visualization?
   A structured method to map data components to graphical attributes

2. What elements are important to include in most data visualizations?
   Answers may vary. Examples: a source dataset; a horizontal axis, a vertical axis; color; geometry/shape

3. In the context of the *grammar of graphics*, what does geometry refer to?
   The shape used to represent data (e.g., bars, lines, points)

4. According to the *grammar of* graphics, what role does color typically play in data visualization?
   It differentiates between data categories or represents values.

Chapter 1: What Is Data Science? (Continued)

Explaining the Concepts

1. Answers may vary. Sample answer: Data are information that can be used to answer a question or motivate a decision.

This idea of data is different from *tidy data*, since *tidy data* is a cleaned version of data. Specifically, *" **Tidy data** is a data structure where each column represents a distinct variable, each row corresponds to a unique observation, and each cell contains only one value"* (Section 1.2).

Excel sheets are a common way to work with data on a day-to-day basis. Excel sheets are similar to tidy data in that they are often in a tabular form. But Excel sheets are not guaranteed to be tidy data since tidy data is typically cleaned data.

3. Answers may vary. Sample answer: When analyzing a large collection of TikTok videos, characteristics, or *variables*, might include: the length of the video, number of views, author of the video, and additional users who are tagged in the video. Each video would be an *observation*. Each column would represent a specific characteristic. Each row would represent a video. A given cell in the table would record the value of the given video's corresponding variable.

When analyzing Amazon reviews, characteristics, or *variables*, might include: the name of the product, an ID for the product (each product should have a unique ID), the text of the review, the rating (an integer from 1 through 5), the name of the author writing the review, a link to the product being reviewed, etc. Each review would be an *observation*. Each column would represent a specific characteristic. Each row would represent a review. A given cell in the table would record the value of the given review's variable.

Identifying the *observations* and *variables* makes the unit of data being studied clear. It will help us perform meaningful analysis, draw conclusions, and create useful visualizations of the data.

5. Answers may vary. Sample answer: Some examples of how one provides data to people, companies, or organizations include: posting to social media, liking and sharing social media posts, website activity, etc. Be more mindful about sharing and reacting to content on social media and disabling cookies on websites that track users.

Some arguments for why this is a good thing include: by providing more information/data we can see better ads. Some arguments why this is a bad thing include: arguments that oversharing our personal information can lead to theft.