Business Statistics - Communicating with Numbers 5th edition Jaggia Solutions Manual

SKU: 9781266273988-SOLUTIONS

Chapter 1. Data and Data Preparation

Solutions

1.
   a. The relevant population consists of all Americans.
   b. The estimates for averages are based on sample data.

2. The value 35 is the estimated average age of the population. It is both costly and time consuming (likely impossible) to take a census of all video game players and compute the actual average age.

3.

   The population is all students enrolled in the accounting class. The value 3.29 represents the population parameter since we are not choosing a sample but drawing results from the actual population.

4.
   a. The population is all recent college graduates with an engineering degree.

   b. No, the average salary is a sample statistic computed from a sample, not the population.

5.
   a. The population is all elderly people. The sample consists of 949 elderly people.
   b. 22% and 17% represent sample statistics.

6.
   a. The relevant population consists of all US citizens born in 2019, excluding non-binary individuals.
   b. The averages are computed from the relevant samples.

7. The resulting data about living accommodations and expenses have a well-defined row-column format, and therefore, are structured. The data are cross-sectional data.
   Note: Data will vary due to the nature of sampling.

8. The data are cross-sectional data.
Note: Data will vary due to the nature of sampling.

9. The data are time series data.
Note: Data will vary depending on the date of retrieval. These data were retrieved July 2019.

Date	Adj. Close Price
7/1/2018	192.55
8/1/2018	195.72
9/1/2018	202.97
10/1/2018	172.33
11/1/2018	176.68
12/1/2018	169.36
1/1/2019	180.90
2/1/2019	182.49
3/1/2019	189.15
4/1/2019	202.28
5/1/2019	188.53
6/1/2019	206.52

10. The front page of the New York Times website is likely to be textual (written reports) with multimedia contents (photographs, etc.). The resulting data are unstructured in that they do not conform to a predefined row-column format.

11. Data on price and fuel economy of small hybrid vehicles can be specified in a predefined row-column format, and therefore, are structured.
Note: Data will vary due to the nature of sampling.

12. The data for Under Armour's annual revenue are structured since they are specified in a well-defined row-column format. The data are time series data.
Note: Data will vary depending on the date of retrieval. These data were retrieved July 2019.

Year	Annual Revenue (in $ millions)
2018	$5,193
2017	$4,989
2016	$4,833
2015	$3,963
2014	$3,084
2013	$2,332

2012	$1,835
2011	$1,473
2010	$1,064
2009	$856

13. The resulting data about online social media usage have a well-defined row-column format, and therefore, are structured. The data are cross-sectional data.
Note: Data will vary due to the nature of sampling.

14.
a. Numerical; discrete
b. Categorical
c. Numerical; continuous

15.
a. Categorical
b. Numerical; continuous
c. Numerical; discrete

16.
a. Nominal
b. Interval
c. Ordinal

17.
a. Ratio
b. Ordinal
c. Nominal

18.
a. Ratio
b. Interval
c. Ratio

19.
a. Nominal scale; the values differ in name.
b.

Major	# of Students
Accounting	4

Economics	4
Finance	4
History	4
Management	5
Psychology	3
Statistics	3
Undecided	3

c. An inspection of the data shows that Management has the highest number of students whereas Psychology, Statistics, and Undecided have the lowest.

20.

a. Nominal
b. Interval. The observations for Year can be ranked, categorized and measured when using this kind of scale. However, there is no true zero point so we cannot calculate meaningful ratios between years.
c. Ratio. This type of scale is the strongest form of measurement. There is a true zero point which allows for the calculation of meaningful ratios between observations.

21.

a. The Purchase variable is categorical whereas Income, Age, and Spending variables are numerical.
b. The measurement scale is nominal for Purchase and ratio for Income, Age, and Spending. Recall, that the nominal and ratio scales represent the least and most sophisticated levels of measurement, respectively.

22.

a. The variable Year is measured on the interval scale because the observations can be ranked, categorized and measured when using this kind of scale. However, there is no true zero point so we cannot calculate meaningful ratios between years.
b. The variable Quarter is measured on the nominal scale, even though it contains numbers. It is the least sophisticated level of measurement because if we are presented with nominal data, all we can do is categorize or group the data.
c. The variable Vacation is measured on the ratio scale. It is the strongest level of measurement because it allows us to categorize and rank the data as well as find meaningful differences between observations. Also, with a true zero point, we can interpret the ratios between observations.

23.

a. 50 observations of x_2 are equal to two.
b. When sorted in ascending order the first observation is $x_1 = 0$ and $x_2 = 1$.
c. When sorted in descending order the first observation is $x_1 = 1$ and $x_2 = 2$.
d. When x_1 is sorted in ascending order and x_2 is sorted in descending order, the first observation is $x_1 = 0$ and $x_2 = 1$.
e. 0 values are missing for x_1 and x_2.

The R Code

```
# Import the data into a data frame (table) and label it myData.
#a
myData <- myData[ , -1]
length(which(myData$x2==2))
#b
sortedData1 <- myData[order(myData$x1, myData$x2), ]
#c
sortedData2 <- myData[order(myData$x1, myData$x2, decreasing = TRUE),]
#d
sortedData3 <- myData[order(myData$x1, -myData$x2), ]
#e
myData[!complete.cases(myData), ]
```

24.

a. 17 observations of x_1 are greater than 30.
b. When sorted in ascending order the first observation is $x_1 = 2$, $x_2 = 540$, and $x_3 = 201$.
c. When x_1 and x_2 are sorted in descending order, and x_3 is sorted in ascending order, the first observation is $x_1 = 52$, $x_2 = 218$, and $x_3 = 154$.
d. 1 value is missing for x_1, 3 values are missing for x_2, and 5 values are missing for x_3.

The R Code

```
# Import the data into a data frame (table) and label it myData.
#a
myData <- myData[ , -1]
length(which(myData$x1>30))
#b
sortedData1 <- myData[order(myData$x1, myData$x2, myData$x3), ]
head(sortedData1)
#c
sortedData2 <- myData[order(myData$x1, myData$x2, decreasing = TRUE, myData$x3), ]
head(sortedData2)
```

```
#d
myData[!complete.cases(myData), ]
```

25.

a. 85 observations of x_4 are less than 3.
b. When sorted in ascending order, the first observation is $x_1 = 0.0121$, $x_2 = 24$, $x_3 = 27$, and $x_4 = 3$.
c. When sorted in descending order, the first observation is $x_1 = 0.9980$, $x_2 = 196$, $x_3 = 13$, and $x_4 = 3$.
d. 0 values are missing for x_1, x_2, x_3, and x_4.
e. The categorical variable x_4 has three categories: 1, 2 and 3. Category 1 has 40 observations, category 2 has 45 observations, and category 3 has 33 observations.

The R Code

```
# Import the data into a data frame (table) and label it myData.
#a
myData <- myData[ , -1]
length(which(myData$x4<3))
#b
sortedData1 <- myData[order(myData$x1, myData$x2, myData$x3, myData$x4), ]
head(sortedData1)
#c
sortedData2 <- myData[order(myData$x1, myData$x2, myData$x3, myData$x4, decreasing
= TRUE), ]
head(sortedData2)
#d
myData[!complete.cases(myData), ]
#e
myData[!complete.cases(myData), ]
length(which(myData$x4==1))
length(which(myData$x4==2))
length(which(myData$x4==3))
```

26.

a. There are 25 observations in the subset.
b. There are 24 observations for $x_4 = 1$ and 28 observations for $x_4 = 0$.

The R Code

```
# Import the data into a data frame (table) and label it myData.
myData$x3 <- as.Date(myData$x3, format = "%m/%d/%Y")
cutoffdate <- as.Date("05/01/1975", format = "%m/%d/%Y")
```

```
subset1 <- myData[myData$x3>=cutoffdate, ]
dim(subset1)
#b
subset2 <- split(myData, myData$x4)
length(subset2$`1`$x1)
length(subset2$`0`$x1)
```

27.

a. There are no missing values
b. 35 observations remain

The R Code

```
# Import the data into a data frame (table) and label it myData.
#a
myData <- myData[, -c(1,5)]
myData[!complete.cases(myData),]
#b
length(which(myData$x2!="Own" & myData$x3 >= 150))
```

28.

a. Variables x_1, x_2 and x_4 have at least one missing value.
b. Observations 13, 18, 21, 25, 35, 40, and 48 have missing values.
c. 45 observations remain after missing values are removed.

The R Code

```
# Import the data into a data frame (table) and label it myData.
#a
myData[!complete.cases(myData),]
#b
which(!complete.cases(myData))
#c
omissionData <- na.omit(myData)
dim(omissionData)
```

29.

a. Minnesota has the highest average writing score of 643. The average math score of Minnesota is 655.
b. The Virgin Islands has the lowest average math score of 445. The average writing score of the Virgin Islands is 490.
c. 13 states reported average math scores above 600.
d. 25 states reported average writing scores below 550.

The R Code

```
# Import the data into a data frame (table) and label it myData.
#a
sortedData1 <- myData[order(myData$Writing, decreasing = TRUE),]
head(sortedData1)
#b
sortedData2 <- myData[order(myData$Math),]
head(sortedData2)
#c
length(which(myData$Math > 600))
#d
length(which(myData$Writing < 550))
```

30.

a. 1 of the 10 highest income earners is married and always exercises.
b. 5 values are missing for Exercise, 2 for Marriage, and 3 for Income.
c. 281 individuals are married, and 134 are not.
d. 69 married individuals always exercise, and 74 unmarried individuals never exercise.
e. 9 individuals are married, exercise sometimes, and earn more than $110,000 per year.

The R Code

```
# Import the data into a data frame (table) and label it myData.
#a
sortedData1 <- myData[order(myData$Income, decreasing = TRUE),]
sortedData1 <- sortedData1[1:10, ]
length(which(sortedData1$Married=="Yes" & sortedData1$Exercise=="Always"))
#b
length(which(is.na(myData$Exercise)))
length(which(is.na(myData$Married)))
length(which(is.na(myData$Income)))
#c
length(which(myData$Married =="Yes"))
length(which(myData$Married == "No"))
#d
length(which(myData$Married == "Yes" & myData$Exercise == "Always"))
length(which(myData$Married == "No" & myData$Exercise == "Never"))
#e
length(which(myData$Married=="Yes" & myData$Exercise=="Sometimes" &
myData$Income > 110000))
```

31.

a. In the data there are 508 males and 382 non-males.
b. $\frac{336}{508} \times 100\% = 66.1417\%$ of males in the data are married, and $\frac{248}{382} \times 100\% = 64.9215\%$ of the non-males in the data are married.
c. Of the 10 individuals with the highest income 7 are married males.
d. The largest income among males is $147,000, the smallest is $35,000. The largest income among non-males is $138,000, the smallest is $22,000.

The R Code

```
# Import the data into a data frame (table) and label it myData.
#a.
length(which(myData$Sex=='Male'))
length(which(myData$Sex=='Non-male'))
#b.
length(which(myData$Sex=='Male' &
myData$Married=='Y'))/length(which(myData$Sex=='Male'))*100
length(which(myData$Sex=='Non-male' &
myData$Married=='Y'))/length(which(myData$Sex=='Non-male'))*100
#c.
sortedData <- myData[order(myData$Income, decreasing = TRUE),]
length(which(sortedData[1:10,]$Sex=='Male' & sortedData[1:10,]$Married=='Y'))
#d.
sortedData2 <- myData[order(myData$Sex, myData$Income, decreasing = FALSE),]
View(sortedData2)
```

32.

a. There are missing values in the variable Travel Plan. All other variables do not have missing values.
b. 300 observations are removed.
c.

College	FoodSpend	Income	TravelPlan
No	1,892.37	77,626	1
No	6,865.77	77,626	1

2 observations remain in the subset.

The R Code

```
# Import the data into a data frame (table) and label it myData.
#a
length(which(is.na(myData$College)))
length(which(is.na(myData$FoodSpend)))
length(which(is.na(myData$Income)))
```

```
length(which(is.na(myData$TravelPlan)))
#b
omissionData <- na.omit(myData)
dim(myData)-dim(omissionData)
#c
subset <- myData[myData$Income>75000 & myData$TravelPlan == 1, ]
View(subset)
```

33.

a. 2018 Population ≥ 5,000,000: 23 observations in this subset.
   2018 Population < 5,000,000: 27 observations in this subset.
b. 9 states are removed.

The R Code

```
# Import the data into a data frame (table) and label it myData.
#a
subset1 <- myData[myData$'2018' >= 5000000, ]
subset2 <- myData[myData$'2018' < 5000000, ]
dim(subset1)
dim(subset2)
#b
subset3 <- subset1[subset1$'2018' <= 10000000, ]
dim(subset1) - dim(subset3)
```

34.

a. The first customer in this list spent $3,362.86 on food.
b. Of the 10 customers who spent the most of travel, 3 were homeowners, and 1 was both a home and car owner.
c. 3 values are missing for the variable OwnHome, 3 values are missing for the variable OwnCar, 0 values are missing for the variable FoodSpend, and 1 value is missing for the variable TravelSpend.
d. 133 of the customers own homes. 88 of the customers own homes but do not own cars.

The R Code

```
# Import the data into a data frame (table) and label it myData.
#a
sortedData1 <- myData[order(myData$OwnHome, myData$OwnCar, myData$TravelSpend,
decreasing = TRUE),]
sortedData1[1,]$FoodSpend
#b
sortedData2 <- myData[order(myData$TravelSpend, decreasing = TRUE),]
```

```
sortedData2 <- sortedData2[1:10,]
length(which(sortedData2$OwnHome=="Yes"))
length(which(sortedData2$OwnHome=="Yes" & sortedData2$OwnCar=="Yes"))
#c
length(which(is.na(myData$OwnHome)))
length(which(is.na(myData$OwnCar)))
length(which(is.na(myData$FoodSpend)))
length(which(is.na(myData$TravelSpend)))
#d
length(which(myData$OwnHome =="Yes"))
length(which(myData$OwnHome == "Yes" & myData$OwnCar == "No"))
```

35.

a. 4 subsets were created: “Dept of HR”, “IT Dept”, “Public Utilities”, and “Sustainability and Environ Dept”.
b. There are 2 missing values in the “IT Dept” Hourly rate variable, all other subsets and variables are complete.

The R Code

```
# Import the data into a data frame (table) and label it myData.
#a
subset1 <- myData[myData$Department == "Public Utilities", ]
subset2 <- myData[myData$Department == "Sustainability & Environ Dept", ]
subset3 <- myData[myData$Department == "IT Dept", ]
subset4 <- myData[myData$Department == "Dept Of HR", ]
#b
subset1[!complete.cases(subset1), ]
subset2[!complete.cases(subset2), ]
subset3[!complete.cases(subset3), ]
subset4[!complete.cases(subset4), ]
```

36.

a. In the data there are 614 females and 616 non-females.
b. $\frac{352}{614} \times 100\% = 57.3290\%$ of the females in the data are admitted.
   $\frac{304}{616} \times 100\% = 49.3506\%$ of the non-females in the data are admitted.
c. 2 of the 10 students with the highest high school GPAs are not female.
d. 6 of the 10 students with the lowest SAT score are female.
e. The highest SAT score of admitted females is 1599, the lowest is 1062. The highest SAT score of admitted non-females is 1600, the lowest is 1055.

The R Code

```
# Import the data into a data frame (table) and label it myData.
#a.
length(which(myData$Female=='Yes'))
length(which(myData$Female=='No'))
#b.
length(which(myData$Decision=='Admit' &
myData$Female=='Yes'))/length(which(myData$Female=='Yes'))*100
length(which(myData$Decision=='Admit' &
myData$Female=='No'))/length(which(myData$Female=='No'))*100
#c.
sortedData <- myData[order(myData$HSGPA, decreasing = TRUE),]
length(which(sortedData[1:10,]$Female == "No"))
#d.
sortedData <- myData[order(myData$SAT),]
length(which(sortedData[1:10,]$Female =="Yes"))
#e.
sortedData <- myData[myData$Female=="Yes" & myData$Decision=="Admit",]
sortedData <- sortedData[order(sortedData$SAT),]
sortedData[dim(sortedData)[1],]$SAT
sortedData[1,]$SAT
sortedData <- myData[myData$Female=="No" & myData$Decision=="Admit",]
sortedData <- sortedData[order(sortedData$SAT),]
sortedData[dim(sortedData)[1],]$SAT
sortedData[1,]$SAT
```

37.

a. Observation 26 is missing for the variable Siblings, observation 13 is missing for the variable Height, observation 47 is missing for the variable Weight, and observations 17 and 51 are missing for the variable Income. All other variables are complete.
b. A total of 5 missing values are removed.

The R Code

```
# Import the data into a data frame (table) and label it myData.
#a
myData[!complete.cases(myData), ]
which(!complete.cases(myData))
#b
omissionData <- na.omit(myData)
dim(myData) - dim(omissionData)
```

38.

a. Yes, all variables have missing values. Observation 20 is missing for Price; 173 is missing for Dividend; 38 is missing for PE; 26 is missing for EPS, 26 and 154 are missing for Lowest, and 46 is missing for Highest.
b. There are 367 complete observations in the subset.
c. 63 observations remain in the data set for which EPS is less than 15.

The R Code

```
# Import the data into a data frame (table) and label it myData.
#a
myData[!complete.cases(myData), ]
which(!complete.cases(myData))
#b
omissionData <- na.omit(myData)
dim(omissionData)
#c
myData1 <- omissionData[which(omissionData$PE<15),]
dim(myData1)
```

Possible Output for Suggested Case Studies

There are many ways to approach these case studies. We provide numerical guidance for one possible scenario around which a report can be written. Instructors may also be interested in Connect multiple-choice exercises (algorithmic) that we have created using the **TechSales_Rep** data. These exercises span multiple chapters, including this one.

Report 1.1

Based on the 2019-2021 rankings, Finland, Denmark, and Iceland are among the happiest countries in the world. Consistent with the past data, Scandinavian countries continue to have favorable rankings compared to other countries. In fact, Finland has occupied the top spot in the Happiness Index Report over the past five years. The report is based on aggregate data on quality of life (e.g., GDP per capita and life expectancy) and a survey of citizens in each country including questions about positive emotions (e.g., enjoyment and laughter) and negative emotions (e.g., anger and sadness). Below is the listing of the top 10 happiest countries according to the 2019-2021 report.

Rank	Country
1	Finland
2	Denmark
3	Iceland
4	Switzerland
5	Netherlands
6	Luxembourg
7	Sweden
8	Norway
9	Israel
10	New Zealand

Report 1.2

First subset the **House_price** data by the college town of Athens (University of Georgia) and the college town of Columbia (University of Missouri). You should obtain 378 observations between these two towns, of which 294 are in Athens and 84 are in Columbia.

- Sale Price is a numerical variable on a ratio scale
- Beds is a numerical variable on a ratio scale
- Baths is a numerical variable on a ratio scale
- Square Footage is a numerical variable on a ratio scale
- Lot size is a numerical variable on a ratio scale
- House Type is a categorical variable on a nominal scale

Students are encouraged to spin a story around the following results.

Percentages	Georgia	Missouri
Sale price $200,000 or less	61.90%	70.24%
Sale price more than $200,000	38.10%	29.76%
Bedrooms 3 or less	61.56%	66.67%
Bedrooms more than 3	38.44%	33.33%
Lot size 10,000 square feet or less	22.79%	36.90%
Lot size more than 10,000 square feet	77.21%	63.10%
Homes built in Year 2000 or earlier	64.97%	51.19%
Homes built after Year 2000	35.03%	48.81%
Single-family homes	98.30%	100.00%

Report 1.3

First subset the **College_Admissions** data by the college to include just Arts & Letters. You should obtain 6964 students.

- Parents' education is a categorical variable on an ordinal scale, represented numerically
- GPA is a numerical variable on a ratio scale
- SAT score is a numerical variable on a ratio scale
- Race is a categorical variable on a nominal scale represented numerically by dummy variables, White and Asian.
- Admitted is a categorical variable on a nominal scale representing the admission decision 'Yes' or No'.

Students are encouraged to spin a story around the following results.

Variable	Percentage
Parent 1 with at least a 4-year college degree	58.73%
Parent 2 with at least a 4-year college degree	55.87%
GPA of 3.5 or more	54.01%
SAT score of 1200 or more	40.34%
White percentage	55.59%
Asian percentage	12.42%
Admitted percentage	24.97%

Report 1.4

First subset the **TechSales_Reps** data to include employees in the software product group with a college degree. You should obtain 9862 observations.

- Feedback is a numerical variable on an interval scale
- Personality type is categorical variable on a nominal scale
- Salary is a numerical variable on a ratio scale
- Net promoter score is a numerical variable on a ratio scale

Students are encouraged to spin a story around the following results.

Percentage	Analyst	Diplomat	Explorer	Sentinel
Feedback more than 3	36.58%	37.29%	37.22%	38.32%
Salary more than 75000	23.28%	48.20%	49.74%	21.38%
NPS more than 8	6.07%	27.57%	27.74%	7.41%
Number of Observations	1,203	3,475	3,673	1,511