GR5702 -- EDAV
Sample Test #1

Name: **UNI:**

Seat #:

You have 75 minutes to complete this test. You are NOT permitted any outside material or assistance to complete this exam.

Instructions:

1. Do not open the test until you are instructed to begin.

2. Enter the ZIP ID that you were assigned in the Student ZipGrade ID section and bubble in the numbers.

3. Bubble in the Key Version on the Answer Sheet. Your Key Version is **A**.

4. Mark your answers both on the test itself and on the bubble answer sheet. (The exam packet will service as a backup in case there's a problem with the bubble sheet.)

5. We will not be able to answer clarification questions during the exam. If you believe that you cannot answer the question as written, provide an explanation in the exam booklet and indicate in the "Notes" section below that we should look at particular question number. For example "see #15".

NOTES

1. A researcher wishes to transform the following time series data so all stock prices are 100 on Day 1. What is the transformed value of Stock B on Day 3?

Day	Stock A	Stock B
1	15	500
2	16	505
3	17	510

(a) 130

(b) 110

(c) 102

(d) 105

2. Displaying a categorical variable in color in a parallel coordinate plot may help identify ___ that wouldn't be visible otherwise.

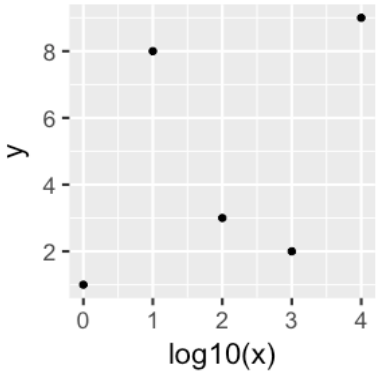
(a) gaps

(b) outliers

(c) clusters

(d) correlations between adjacent variables

3. What is the range of x values shown in the scatterplot below?



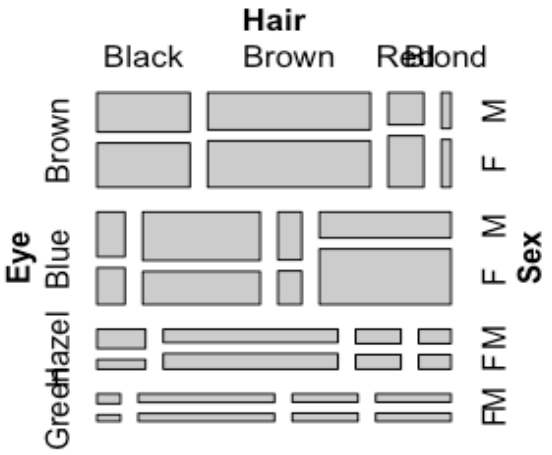
(a) (1000, 9000)

(b) (1, 10000)

(c) (0, 40)

(d) (10, 1000)

4. Indicate the order in which the variables were split in the construction of the following mosaic chart:



(a) 1. Hair 2. Eye 3. Sex

(b) 1. Hair 2. Sex 3. Eye

(c) 1. Eye 2. Sex 3. Hair

(d) 1. Eye 2. Hair 3. Sex

5. Splines are particularly useful for parallel coordinate plots in cases in which ___ .

(a) each variable is approximately uniformly distributed

(b) adjacent variables are not correlated

(c) the units are very different for different variables

(d) there are many repeated values

6. If you were to draw a boxplot of the dataset listed below (both the values and the five number summary are provided), which of the values, if any, would be considered outliers?

   Values: -60, -49, -22, -18, -5, 22, 25, 25, 25, 25, 25, 25, 26, 34, 47, 52, 64, 68, 90, 143, 150

   Five number summary: -60, 22, 25, 52, 150

   (a) -60, -49, 90, 143, 150

   (b) No outliers

   (c) -60, 143, 150

   (d) -60, -49, 143, 150

7. The common baseline(s) for judging length in a diverging stacked bar chart using standard Likert data categories is(are):

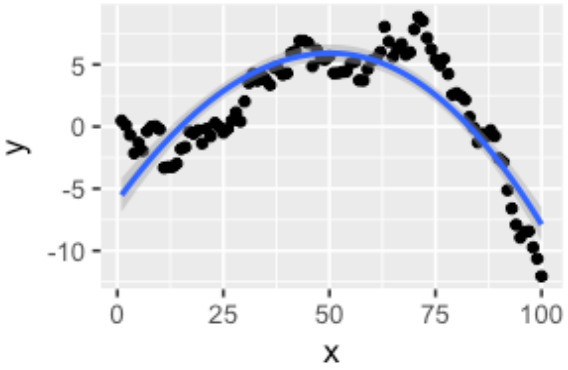
   (a) the middle or neutral category

   (b) the strongly disagree category

   (c) both strongly agree and strongly disagree categories

   (d) the strongly agree category

8. The following is an example of _____ to the data.



   (a) overfitting a loess smoother

   (b) overfitting a linear model

   (c) underfitting a loess smoother

   (d) properly fitting a loess smoother

9. Which of the following is best for identifying modes in the distribution of a continuous variable?

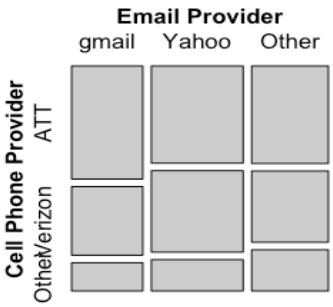
   (a) density curve

   (b) scatterplot

   (c) boxplot

   (d) mosaic plot

10. A random sample of 175 Cal Poly State University students was selected, and both the email service provider and cell phone provider were determined for each one, resulting in the accompanying mosaic plot:

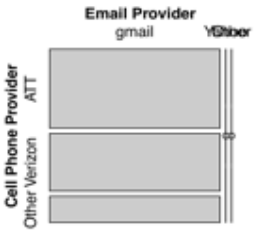


Given this data, which of the following shows no association between email service provide and cell phone provider? (That is, the expected values for the null hypothesis in a chi square test.)

A)

B)

C)



D)

11. Which of the following is best for identifying outliers in the distribution of a single continuous variable?

   (a) density curve

   (b) boxplot

   (c) scatterplot

   (d) mosaic plot

12. Topcoding in a plot is problematic if

    (a) there's not a lot of data in the topcoded category

    (b) the data is ordinal

    (c) there's a lot of data in the topcoded category

    (d) the topcoded category appears on the x-axis

13. Which of the following are appropriate for visualising single dimensional categorical data?

    (a) boxplot and Cleveland dot plot

    (b) histogram and bar chart

    (c) bar chart and Cleveland dot plot

    (d) boxplot and bar chart

14. Which of the following cannot be inferred from the mosaic plot, which shows treatment results (marked = "a lot", some, none) for a study in which some patients were treated and others received a placebo?

(a) Men were more likely than women to improve with treatment.

(b) No men who received the placebo reported some improvement.

(c) There were more women in the study than men.

(d) Women were more likely than men to improve with the placebo.

15. Which of the following would be the best strategy for identifying a monthly effect in time series data? Suppose you have plotted daily values for a five year period.

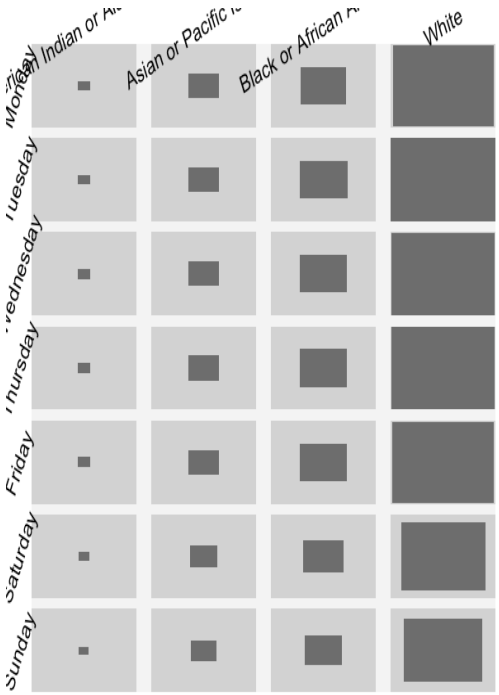
    (a) Add a smoothing curve

    (b) Draw a separate line for each of the twelve months

    (c) Facet on year

    (d) Plot the monthly average over time

16. The following is a fluctuation diagram showing birth day of week ("day") vs. race of mother ("race") for U.S. births in 2015. Which of the following can be inferred from the diagram?



    (a) strong association between race and day

    (b) deterministic relationship between race and day

    (c) no association between race and day

    (d) moderate association between race and day

17. A loess smoother:

    (a) fits a sinusoidal model to the data

    (b) does not assume any model

    (c) fits a log model to the data

    (d) fits a polynomial model to the data

18. Which of the following is tidy data?

Decline in Male Births

A data frame with 45 observations on the following 4 variables.

`Year` year of observation

`Denmark` male birth rate of Denmark for given year

`Netherlands` male birth rate of The Netherlands for given year

`Canada` male birth rate of Canada for given year

Election

A data frame with 67 observations on the following 3 variables.

`County` a character vector indicating the county

`Buchanan2000` votes cast for P. Buchanan

`Bush2000` votes cast for G.W. Bush

Corn

A data frame with 38 observations on the following 3 variables.

`Year` year of observation (1890–1927)

`Yield` average corn yield for the six states (in bu/acre)

`Rainfall` average rainfall in the six states (in in/year)

Tires

A data frame with 8 observations on the following 4 variables.

`SpeedCat` a numerical code corresponding to 4 categories of speed (in miles per hour), with 1 = "0-40", 2 = "41-55", 3 = "56-65" and 4 = ">65"

`Make` a factor with levels "Ford" and "Other"

`Other` cause of accident was other than tire-related (frequency count)

`Tire` cause of accident was tire-related (frequency count)

(a) Corn

(b) Decline in Male Births

(c) Election

(d) Tires

19. Which of the following is NOT appropriate for studying the relationship between two quantitative variables?

    (a) performing a regression analysis

    (b) finding the correlation coefficient

    (c) drawing a scatterplot

    (d) drawing a mosaic plot

20. Based on the summary of the cross-section data for 675 14-year old children born between 1980 and 1988, which of the following is NOT true?

```
## Observations: 675
## Variables: 12
## $ school      <fct> Gymnasium, Gymnasium, Gymnasium, Gymnasium, Realsc...
## $ birthyear   <int> 1981, 1981, 1980, 1984, 1982, 1980, 1986, 1986, 19...
## $ gender      <fct> female, female, female, female, male, female, fema...
## $ kids        <int> 2, 2, 3, 1, 4, 3, 3, 1, 4, 2, 2, 2, 5, 5, 5, 2, 1,...
## $ parity      <int> 2, 2, 3, 1, 4, 1, 3, 1, 4, 2, 1, 1, 3, 2, 1, 2, 1,...
## $ income      <dbl> 35160.11, 65748.35, 120962.36, 60100.57, 34828.95,...
## $ size        <dbl> 4, 3, 3, 3, 4, 5, 3, 3, 5, 4, 4, 4, 7, 7, 7, 4, 3,...
## $ state       <fct> Berlin, Berlin, Berlin, Berlin, Berlin, Berlin, Be...
## $ marital     <fct> married, married, married, married, divorced, marr...
## $ meducation  <dbl> 14.5, 10.5, 12.0, 10.5, 10.0, 15.0, 15.0, 13.0, 15...
## $ memployment <fct> none, parttime, parttime, parttime, fulltime, part...
## $ year        <dbl> 1995, 1995, 1994, 1998, 1996, 1994, 2000, 2000, 19...
```

    (a) `gender` is a categorical variable

    (b) `state` is an ordinal variable

    (c) `parity` is a discrete variable

    (d) `memployment` is a nominal variable

21. Scatterplots are not particulary useful for identifying which of the following?

    (a) outliers

    (b) positive or negative correlations between variables

    (c) multimodality in each variable

    (d) associations between variables

22. The philosophy of the grammar of graphics is to:

    (a) create a language for describing the underlying structure of data visualizations

    (b) create a comprehensive list of all chart types

    (c) devise a system for choosing the best graph type based on the data

    (d) devise an algorithm for translating the main point of data visualizations into sentence structure

23. When plotting the density curve of a continous variable, if the curve is too ragged (lots of little ups and downs) for you to understand the global features of the data, what should you do?

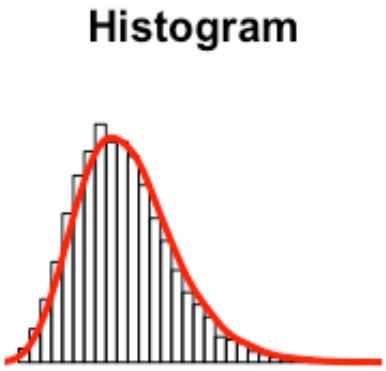
    (a) shrink the plotting range of the data

    (b) decrease the binwidth

    (c) extend the plotting range of the data

    (d) increase the binwidth

24. Based on the following histogram, which of the following best describes the shape of the distribution?



    (a) right skewed

    (b) left skewed

    (c) multimodal

    (d) symmetric

25. Based on the results of Cleveland's research, rank the following visual perception tasks in order from the easiest to perform successfully to the most difficult:

    Color [1]

    Position along a common scale [2]

    Angle [3]

    Position along identical, non-aligned scales [4]

    (a) 2, 4, 1, 3

    (b) 2, 4, 3, 1

    (c) 2, 1, 4, 3

    (d) 2, 1, 3, 4

26. Which is the best graphic option for testing normality of a continuous variable?

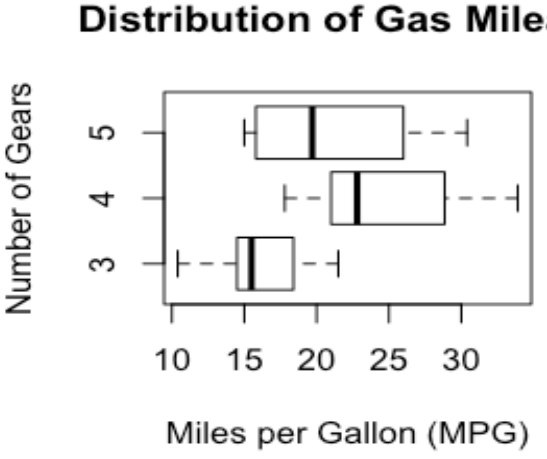
    (a) boxplot

    (b) stem and leaf plot

    (c) Q-Q plot

    (d) bar chart

27. The gas mileages of different types of cars (number of forward gears is 3,4 or 5) are summarized in the boxplots below. Which of the following statements is NOT true?



(a) About 25% cars with 4 gears are more than 29 MPG

(b) The median MPG cars with 3 gears is about 16

(c) The cars with 4 gears have less variability in MPG than cars with 3 gears

(d) About 50% cars with 5 gears are between 16 and 27 MPG

28. Ordinal data should be ordered in a bar chart:

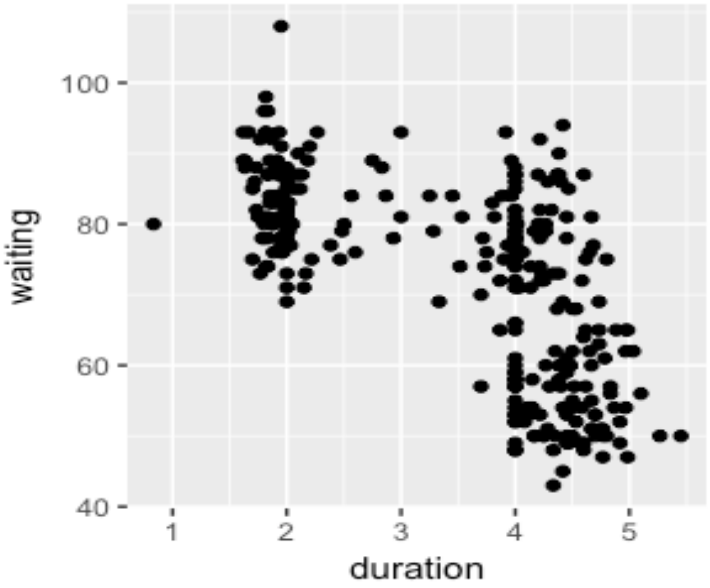
    (a) according to the natural order of the categories

    (b) in increasing order of frequency

    (c) in alphabetical order

    (d) in decreasing order of frequency

29. The following scatterplot shows the relationship between the waiting time to the next eruption and the duration of the current eruption for the Old Faithful geyser in Yellowstone National Park.



Which of the following statements is NOT correct?

    (a) There are a few outliers.

    (b) The values for the eruption durations appear to be rounded.

    (c) There appear to be three clusters.

    (d) A short duration time is associated with a short waiting time until the next eruption.

30. Which of the histograms below could be drawn from the same data as that used to draw the boxplot?