Exercise 3.2

Kyle Ramirez

12/19/2021

```
{r setup, include=FALSE} knitr::opts_chunk$set(echo = TRUE)
```

1. For this exercise, you will use the following dataset, 2014 American Community Survey. This data is maintained by the US Census Bureau and are designed to show how communities are changing. Through asking questions of a sample of the population, it produces national data on more than 35 categories of information, such as education, income, housing, and employment. For this assignment, you will need to load and activate the ggplot2 package. For this deliverable, you should provide the following:

1. What are the elements in your data (including the categories and data types)?

2. Please provide the output from the following functions: str(); nrow(); ncol()

3. Create a Histogram of the HSDegree variable using the ggplot2 package.

1. Set a bin size for the Histogram.

2. Include a Title and appropriate X/Y axis labels on your Histogram Plot.

4. Answer the following questions based on the Histogram produced:

1. Based on what you see in this histogram, is the data distribution unimodal?

2. Is it approximately symmetrical?

3. Is it approximately bell-shaped?

4. Is it approximately normal?

5. If not normal, is the distribution skewed? If so, in which direction?

6. Include a normal curve to the Histogram that you plotted.

7. Explain whether a normal distribution can accurately be used as a model for this data.

Set the working directory to the root of your DSC 520 directory

setwd("/Users/Kyle/Documents/GitHub/KR/Ramirez_Kyle_DSC510/dsc520/data")

Examine the structure of `surveyData_df` using 'summary(), str(), nrow(), ncol()'

surveyData_df <- read.csv('acs-14-1yr-s0201.csv', stringsAsFactors = FALSE) summary(surveyData_df) str(surveyData_df) nrow(surveyData_df) ncol(surveyData_df)

Create a histogram of the `HSDegree` variable using `geom_histogram()`

Use 30 bins, add title:, xlab(), ylab()

ggplot(surveyData_df, aes(HSDegree)) + geom_histogram(bins = 30) + ggtitle('High School Degree Distribution') + xlab('Percentage of People Who Completed HS Degrees') + ylab('Counts')

Question 1: Based on what you see in this histogram, is the data distribution unimodal?

Answer: No

Question 2: Is it approximately symmetrical?

Answer: No

Question 3: Is it approximately bell-shaped?

Answer: Yes

Question 4: Is it approximately normal?

Answer: No

Question 5: If not normal, is the distribution skewed? If so, in which direction?

Answer: Skewed to the left.

Question 6: Include a normal curve to the Histogram that you plotted.

Answer: This is a bell shaped curve skewed to the left.

x <- seq(60, 120, by = .1) y <- dnorm(x, mean = 0, sd = 15)

ggplot(surveyData_df, aes(HSDegree)) + geom_histogram(bins = 30) + ggtitle('High School Degree Distribution') + xlab('Percentage of People Who Completed HS Degrees') + ylab('Counts') + lines(x, y) +

Question 7: Explain whether a normal distribution can accurately be used as a model for this data.

Answer: This is not a normal distribution as the graph is skewed to the left.

Create a Probability Plot of the HSDegree variable.

```
set.seed(0)
ggplot(surveyData_df, aes(HSDegree, RacesReported)) + geom_qq() + geom_qq_line()
qplot(x = HSDegree)
```