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

Data Category Data Types

Id Varchar

Id2 Number

Geography Varchar

PopGroupID Boolean

POPGROUP.display-label Varchar

RacesReported NUmber

HSDegree Decimal

BachDegree Decimal

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

Please Refer git hub links

Code :

library (pastecs)

library(ggplot2)

theme\_set(theme\_minimal())

amcomsurvey <- read.csv("data/acs-14-1yr-s0201.csv")

#str(),nrow(),ncol()

str(amcomsurvey)

nrow(amcomsurvey)

ncol(amcomsurvey)

Ouptut

A picture containing text, font, measuring stick

Description automatically generated

1. 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.

Code :

ggplot(amcomsurvey,aes(HSDegree)) + geom\_histogram(bins=30)

#Histogram with bin size + titles

ggplot(amcomsurvey,aes(HSDegree)) + geom\_histogram(bins=30) + ggtitle('HSDegree Histogram') + xlab('High School Degree') + ylab('Points')

Output:

A picture containing diagram, screenshot, plot, line

Description automatically generated

A graph of a high school degree

Description automatically generated with low confidence

1. Answer the following questions based on the Histogram produced:
   1. Based on what you see in this histogram, is the data distribution unimodal?

Ans: As there is only one peak, it is a unimodal distribution.

* 1. Is it approximately symmetrical?

Ans : The distribution doesn’t seem to be a symmetrical.

* 1. Is it approximately bell-shaped?

Ans : The Distribution doesn’t seem to be a bell shaped.

* 1. Is it approximately normal?

ns: It is not a normal distribution exactly, but we can say it is approximately normal

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

Ans: It is a negatively skewed distribution because the tail is pointing towards the lower values of HSDegree.

* 1. Include a normal curve to the Histogram that you plotted.

Code :

ggplot(amcomsurvey,aes(HSDegree)) + geom\_histogram(aes(y = ..density..), bins = 30) + geom\_function(fun = dnorm, args = list(mean = mean(amcomsurvey$HSDegree), sd = sd(amcomsurvey$HSDegree)), size=3, colour="Red")

Output:

A graph of a high school degree

Description automatically generated with low confidence

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

Ans: normal probability model can be used even if the distribution of the continuous outcome is not perfectly symmetrical; it just has to be reasonably close to a normal

1. Create a Probability Plot of the HSDegree variable.

Code :

qqnorm(amcomsurvey$HSDegree)

qqline(amcomsurvey$HSDegree)

Output:

A picture containing text, line, diagram, plot

Description automatically generated

1. Answer the following questions based on the Probability Plot:
   1. Based on what you see in this probability plot, is the distribution approximately normal? Explain how you know.

Ans: The plotted points bend down and to the right of the normal line that indicates a long tail to the left. It indicates that the distribution is not normal.

* 1. If not normal, is the distribution skewed? If so, in which direction? Explain how you know.

Ans: The distribution is left skewed. The plotted points bend down and to the right of the normal line that indicates a long tail to the left.

1. Now that you have looked at this data visually for normality, you will now quantify normality with numbers using the stat.desc() function. Include a screen capture of the results produced.

Ans: Please find screen capture below.

A screenshot of a computer

Description automatically generated with low confidence

1. In several sentences provide an explanation of the result produced for skew, kurtosis, and z-scores. In addition, explain how a change in the sample size may change your explanation?

**Skew:**

Skew is the lack of symmetry in the data. As we can see the plot is not exactly normally distributed and it’s tail is pointing towards the lower numbers, it is negatively skewed data.

Kurtosis:

Kurtosis is the pointiness of the data. The plot seems to be heavily tailed; it is leptokurtosis.

**Z Scores:**

Z score is the relationship of a score with the mean of the score and the. To Calculate Z score of X the formula is

Z = x – mean / standard deviation.

From the given dataset lets find the Z score of value 88.1

Mean of HSDegree Values is 87.63235

Standard deviation of HSDegree is 5.117941

Z score = (88.1 -87.63) /5.11

= 0.09

It means that the value 88.1 is 0.09 Standard Deviation away from the mean. It implies that the value is pretty much close to the mean.

**How The Sample Size impacts the Skew, Kurtosis and Z score**.

I am not sure how to explain this or do we need to explain with an example. Here is what I came up with.

The Skew, Kurtosis and Z score are mainly dependent on the mean and standard deviation. If we take a smaller sample size it will be very misleading as the mean value will not be far from the mean of the sample. Bigger sample size leads to the mean value which is close to the mean of the population and increases the possibility of model to be a good fit.