# Assignment: ASSIGNMENT 3\_2

# Name: Kalaikkovan, Vasanthakumar

# Date: 2010-03-30

## Load the ggplot2 package

**library(ggplot2)**

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

**setwd("E:/Repos/StatisticsR/DSC520-Statistics/")**

## Load the `data/r4ds/heights.csv` to

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

**class(df)**

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

## Categorical, numerical and binary

**names(df)**

**typeof(df$Id)**

**class(df$Id)**

# Id - Categorical

# Id2 - numerical

# Geography - Categorical

# PopGroupID - binary

# POPGROUP.display.label - Categorical

# RacesReported - numerical

# HSDegree - numerical

# BachDegree - numerical

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

**str(df)**

# Answer

# 'data.frame': 136 obs. of 8 variables:

# $ Id : chr "0500000US01073" "0500000US04013" "0500000US04019" "0500000US06001" ...

# $ Id2 : int 1073 4013 4019 6001 6013 6019 6029 6037 6059 6065 ...

# $ Geography : chr "Jefferson County, Alabama" "Maricopa County, Arizona" "Pima County, Arizona" "Alameda County, California" ...

# $ PopGroupID : int 1 1 1 1 1 1 1 1 1 1 ...

# $ POPGROUP.display.label: chr "Total population" "Total population" "Total population" "Total population" ...

# $ RacesReported : int 660793 4087191 1004516 1610921 1111339 965974 874589 10116705 3145515 2329271 ...

# $ HSDegree : num 89.1 86.8 88 86.9 88.8 73.6 74.5 77.5 84.6 80.6 ...

# $ BachDegree : num 30.5 30.2 30.8 42.8 39.7 19.7 15.4 30.3 38 20.7 ...

**nrow(df)**

## Answer 136

**ncol(df)**

## Answer 8

iii. Create a Histogram of the HSDegree variable using the ggplot2 package. Set a bin size for the Histogram. Include a Title and appropriate X/Y axis labels on your Histogram Plot.

**x <-df$HSDegree**

**hist(x, breaks = 40, xlab = "Percentage of HS Degree", ylab = "Frequency", main = "Percentage of HS Degree Frequency Distribution")**

Chart, histogram

Description automatically generated

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

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

**Answer - It is a unimodal distribution because it has only one peak.**

1. Is it approximately symmetrical?

**Answer - No its skewed**

1. Is it approximately bell-shaped?

**Answer - No its skewed**

1. Is it approximately normal?

**Answer - No its skewed**

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

**Answer - Yes, its skewed left (negative)**

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

**h <- hist(x, breaks = 10, density = 10, col = "lightgray", xlab = "Percentage of HS Degree", ylab = "Frequency", main = "Percentage of HS Degree Frequency Distribution")**

**xfit <- seq(min(x), max(x), length = 40)**

**yfit <- dnorm(xfit, mean = mean(x), sd = sd(x))**

**yfit <- yfit \* diff(h$mids[1:2]) \* length(x)**

**lines(xfit, yfit, col = "black", lwd = 2)**

Chart

Description automatically generated

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

**Answer - No we cannot use normal distribution in this model because it is left (negative) skewed.**

v. Create a Probability Plot of the HSDegree variable.

**install.packages("qqplotr")**

**library(qqplotr)**

**ggplot(mapping = aes(sample = df$HSDegree)) + stat\_qq\_point(size = 2)**

Chart, line chart, scatter chart

Description automatically generated

vi. 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.

**Answer - This probability plot is not approximately normal because it skewed.**

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

**Answer - The plotted points are bend down and to the right of the normal line that indicates a long tail to the left. So, its left (negative) skewed.**

vii. 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.

**install.packages("pastecs")**

**library(pastecs)**

**stat.desc(df)**

A picture containing text

Description automatically generated

viii. 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?

**Skewness- It is a measure of the asymmetry in the distribution, and it’s skewed in the left side. So, it is a left skew distribution.**

**kurtosis - is a measure of peakiness of a distribution.**

**z-score - A z-score could be obtained by dividing the skew values or excess kurtosis by their standard errors.**

**Transformation methods like following can be used on the sample to change the distribution.**

**square-root for moderate skew:**

**sqrt(max(x+1) - x)**

**log for greater skew:**

**log10(max(x+1) - x)**

**inverse for severe skew:**

**1/(max(x+1) - x)**