assignment 2.1 NelsonRachel

Rachel Nelson

6/13/2020

Loads and activate the ggplot2 and pastecs packages. :

```
library(ggplot2)
library(pastecs)
library(latexpdf)
setwd("C:/Users/Rachel/Desktop/College/DSC520/dsc520-master")
acs_df <- read.csv("data/acs-14-1yr-s0201.csv")
```

1. What are the elements in your data (including the categories and data types)? The elements include ID=factor, ID2=int, Geography=factor, PhotoGroupID=int, PopGroup=factor, Races Reported=int, HSDegree=num, BachDegree=nium

```
##               Id           Id2                                  Geography
##  0500000US01073:  1   Min.   : 1073    Alameda County, California    :  1
##  0500000US04013:  1   1st Qu.:12082    Allegheny County, Pennsylvania:  1
##  0500000US04019:  1   Median :26112    Anne Arundel County, Maryland :  1
##  0500000US06001:  1   Mean   :26833    Arapahoe County, Colorado     :  1
##  0500000US06013:  1   3rd Qu.:39123    Baltimore city, Maryland      :  1
##  0500000US06019:  1   Max.   :55079    Baltimore County, Maryland    :  1
##  (Other)       :130                    (Other)                       :130
##    PopGroupID      POPGROUP.display.label RacesReported          HSDegree
##  Min.   :1    Total population:136         Min.   :  500292   Min.   :62.20
##  1st Qu.:1                                 1st Qu.:  631380   1st Qu.:85.50
##  Median :1                                 Median :  832708   Median :88.70
##  Mean   :1                                 Mean   : 1144401   Mean   :87.63
##  3rd Qu.:1                                 3rd Qu.: 1216862   3rd Qu.:90.75
##  Max.   :1                                 Max.   :10116705   Max.   :95.50
##
##    BachDegree
##  Min.   :15.40
##  1st Qu.:29.65
##  Median :34.10
##  Mean   :35.46
##  3rd Qu.:42.08
##  Max.   :60.30
##
```

```
## $names
## [1] "Id"                     "Id2"                    "Geography"
## [4] "PopGroupID"             "POPGROUP.display.label" "RacesReported"
## [7] "HSDegree"                "BachDegree"
##
## $class
## [1] "data.frame"
##
## $row.names
##   [1]   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18
##  [19]  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36
##  [37]  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54
##  [55]  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72
##  [73]  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90
##  [91]  91  92  93  94  95  96  97  98  99 100 101 102 103 104 105 106 107 108
## [109] 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126
## [127] 127 128 129 130 131 132 133 134 135 136
```

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

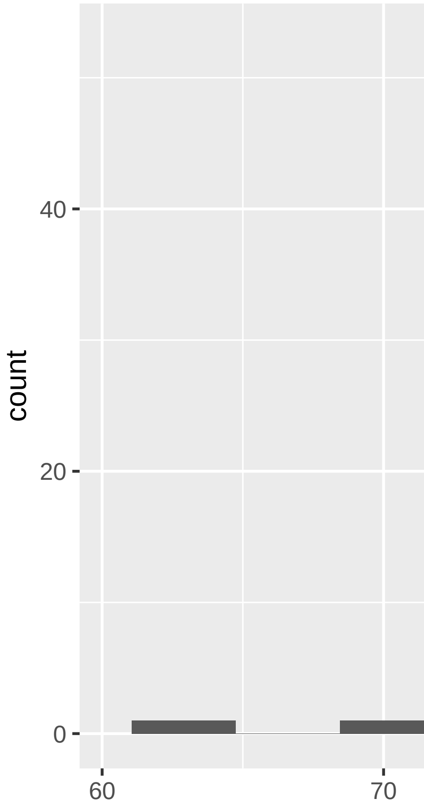
```
## 'data.frame':    136 obs. of  8 variables:
##  $ Id                    : Factor w/ 136 levels "0500000US01073",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ Id2                   : int  1073 4013 4019 6001 6013 6019 6029 6037 6059 6065 ...
##  $ Geography             : Factor w/ 136 levels "Alameda County, California",..: 56 70 98 1 20 43 62
##  $ PopGroupID            : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ POPGROUP.display.label: Factor w/ 1 level "Total population": 1 1 1 1 1 1 1 1 1 1 ...
##  $ RacesReported         : int  660793 4087191 1004516 1610921 1111339 965974 874589 10116705 3145515
##  $ 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 ...
```

```
## [1] 136
```

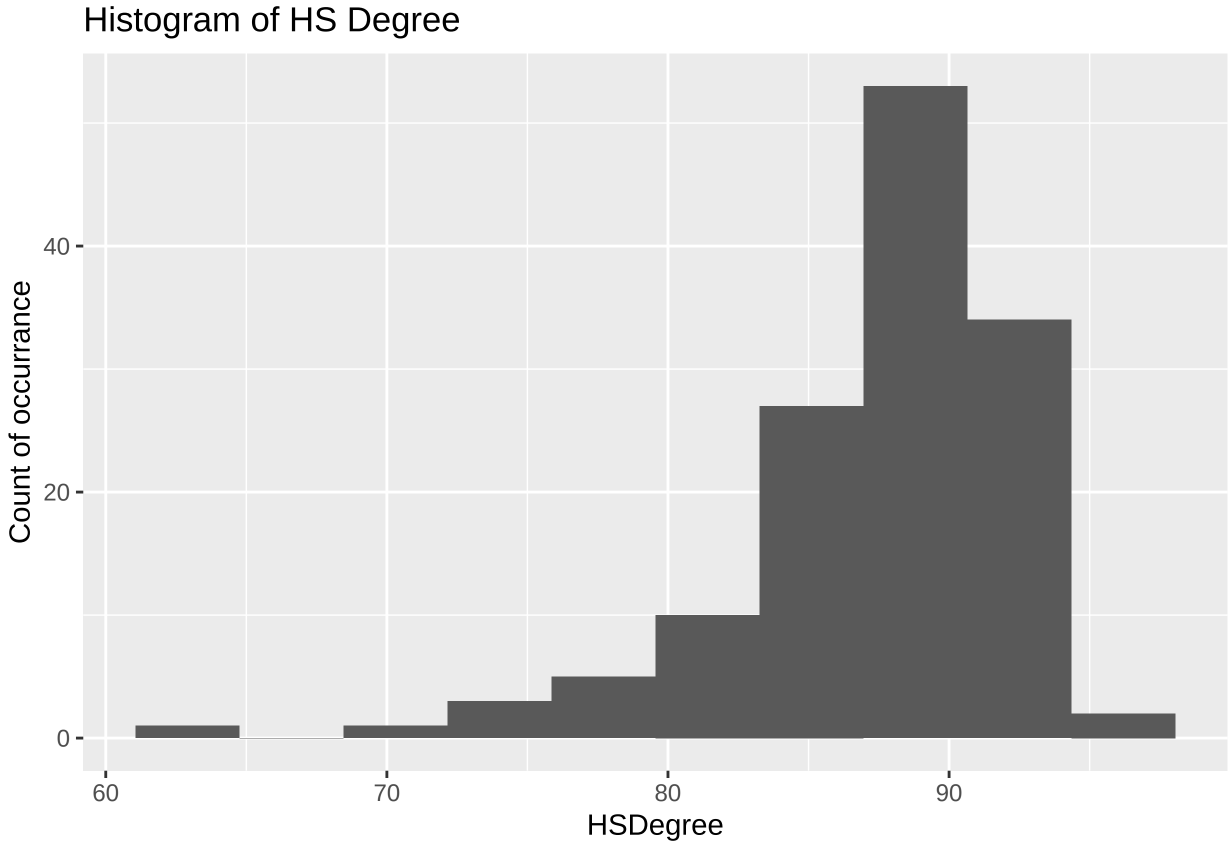
```
## [1] 8
```

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

a. Set a bin size for the Histogram.



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



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

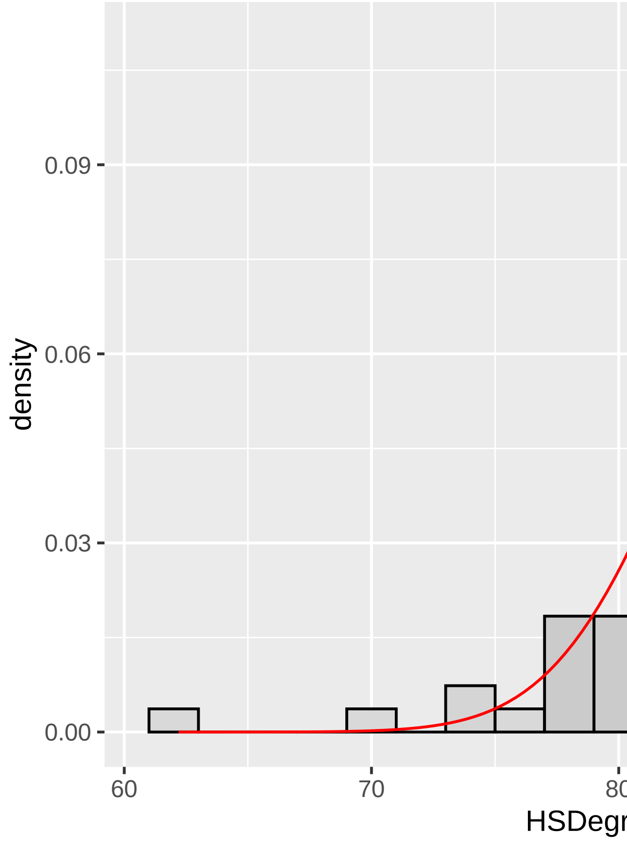
a. Based on what you see in this histogram, is the data distribution unimodal? Yes, the distribution is unimodal as it has one clear peak

b. Is it approximately symmetrical? No, the data is skewed to the left

c. Is it approximately bell-shaped? No

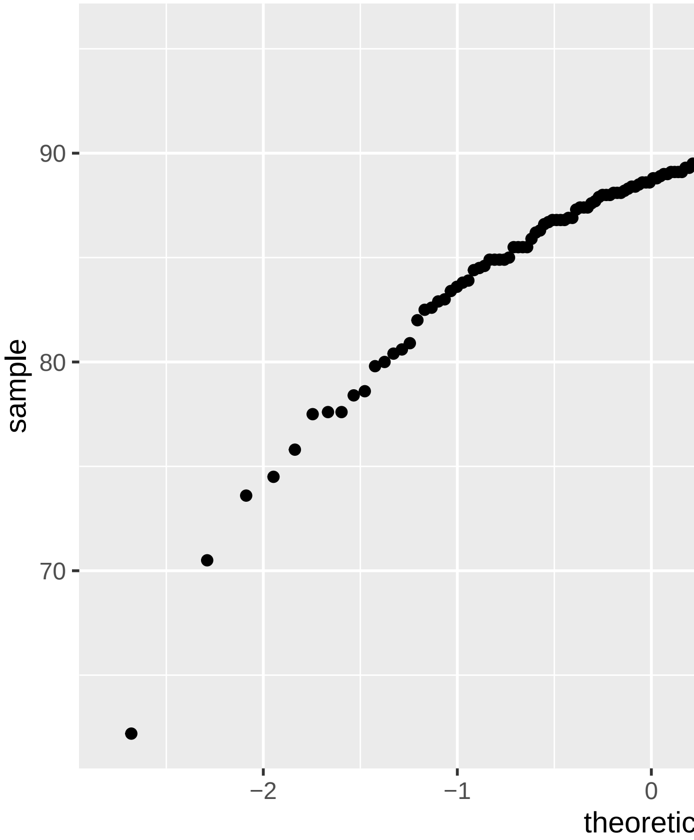
d. Is it approximately normal? No, the lack of it being bell-shaped and symmetrical are also indicators that the data is not likely normal

e. If not normal, is the distribution skewed? If so, in which direction? Yes, the data is skewed to the left



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

g. Explain whether a normal distribution can accurately be used as a model for this data. No, because the data does not have the characteristics of normal data (99.9% of the data will not be within six sigma (or threee standard deviations either way)) since the data is skewed left



5. Create a Probability Plot of the HSDegree variable.

6. Answer the following questions based on the Probability Plot:

a. Based on what you see in this probability plot, is the distribution approximately normal? Explain how you know. No, if the data was normal the line would be linear. Because of the curvature, you can tell that the data is nor normal.

b. If not normal, is the distribution skewed? If so, in which direction? Explain how you know. Yes, the distribution is skewed to the left. You can tell by the downward curvature of the line.

7. 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.

```
##           Id          Id2 Geography PopGroupID POPGROUP.display.label
## nbr.val   NA 1.360000e+02        NA        136                     NA
## nbr.null  NA 0.000000e+00        NA          0                     NA
## nbr.na    NA 0.000000e+00        NA          0                     NA
## min       NA 1.073000e+03        NA          1                     NA
## max       NA 5.507900e+04        NA          1                     NA
## range     NA 5.400600e+04        NA          0                     NA
## sum       NA 3.649306e+06        NA        136                     NA
## median    NA 2.611200e+04        NA          1                     NA
## mean      NA 2.683313e+04        NA          1                     NA
## SE.mean   NA 1.323036e+03        NA          0                     NA
## CI.mean  NA 2.616557e+03          NA          0                     NA
## var      NA 2.380576e+08          NA          0                     NA
## std.dev  NA 1.542911e+04          NA          0                     NA
## coef.var NA 5.750024e-01          NA          0                     NA
##          RacesReported     HSDegree   BachDegree
## nbr.val   1.360000e+02 1.360000e+02  136.0000000
## nbr.null  0.000000e+00 0.000000e+00    0.0000000
## nbr.na    0.000000e+00 0.000000e+00    0.0000000
## min       5.002920e+05 6.220000e+01   15.4000000
## max       1.011671e+07 9.550000e+01   60.3000000
## range     9.616413e+06 3.330000e+01   44.9000000
## sum       1.556385e+08 1.191800e+04 4822.7000000
## median    8.327075e+05 8.870000e+01   34.1000000
## mean      1.144401e+06 8.763235e+01   35.4610294
## SE.mean   9.351028e+04 4.388598e-01    0.8154527
## CI.mean   1.849346e+05 8.679296e-01    1.6127146
## var       1.189207e+12 2.619332e+01   90.4349886
## std.dev   1.090508e+06 5.117941e+00    9.5097313
## coef.var  9.529072e-01 5.840241e-02    0.2681741
```

```
##        median          mean       SE.mean  CI.mean.0.95           var
##  8.870000e+01  8.763235e+01  4.388598e-01  8.679296e-01  2.619332e+01
##       std.dev      coef.var      skewness      skew.2SE      kurtosis
##  5.117941e+00  5.840241e-02 -1.674767e+00 -4.030254e+00  4.352856e+00
##      kurt.2SE    normtest.W    normtest.p
##  5.273885e+00  8.773635e-01  3.193634e-09
```

8. 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 measures the asymmetry of the data and kurtosis measure the peakedness of the data. For skew, a score of 0 represents a perfect normal distribution. In this case, the skew is -1.67, the negative number indicates the data is skewed to the left For kutorsis, a value of 0 represents a perfect normal distribution. In this case, the kurtosis is 4.35, which indicates the peakedness of the data is not considered normal The Z score For medium-sized samples, you can reject the null hypothesis at absolute z-value over 3.29,and determine the data is considered non-normal. The more samples, the greater probability of rejecting that the values come from a normal distribution because it becoems more sensitive to small deviations within the data

tinytex::install_tinytex() tinytex:::is_tinytex()