**The Github repo link for the linked R file:**

https://github.com/tang-xin72/DSC520/tree/main/completed/assignment03

A screenshot of a computer

Description automatically generated with medium confidence



**American Community Survey Exercise:**

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

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

$ Id : chr

$ Id2 : integer

$ Geography : chr

$ PopGroupID : integer

$ POPGROUP.display.label: chr

$ RacesReported : integer

$ HSDegree : number

$ BachDegree : number

A picture containing diagram, screenshot, plot, line

Description automatically generated

**Figure 1: high school degree holder percentage per population in different counties**

1. **Histogram observation**

* This data distribution is unimodal.
* The distribution is not symmetrical.
* The distribution is close to a bell shape but has multiple high peaks with concentrated bell shape.
* The distribution is not normal.
* The distribution is negatively skewed.
* A normal distribution is not preferred to model this data.
* A picture containing screenshot, line, plot, diagram

  Description automatically generated

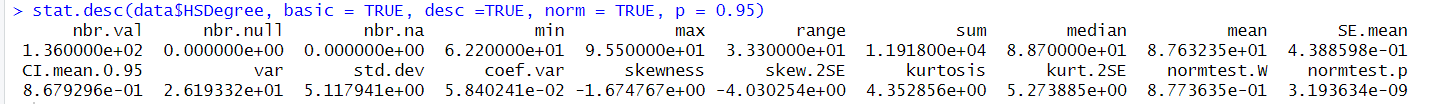
A picture containing line, diagram, plot, screenshot

Description automatically generated

**Figure 2: probability chart: high school degree holder percentage per population in different counties**

**probability observation**

* + **This distribution is not normal, since it has peak not centered.**
  + **It is negatively skewed toward right side. The left side has long tail but right side has sharp and short decline.**



**Summary**:

* The skewness = -1.67, since it is less than \_1, it is highly skewed and skewed toward left. Which is proven by chart above.
* The kurtosis = 4.35, a normal distribution should get 3. So it have thin bell.
* Since the kurt.2SE = 5.27 and skew.2SE = -4.03. so the skew is unlikely by chance. It is a real skewness.
* Since Z score is normalized, this analysis works for small samples size, if the sample size is large, it won’t be accurate and need to look at the actual distribution shape.