DSC520

Assignment Week 3

Zemelak Goraga

2023-12-16"

In [5]:
```
# Set the working directory
# setwd("labs/R101")

# Read the CSV file into a data frame as df
df <- read.csv("df.csv")

# Display the first few rows of the dataset
head(df)
```

A data.frame: 6 × 8

	Id	Id2	Geography	PopGroupID	POPGROUP.display.label	RacesRepor
	<fct>	<int>	<fct>	<int>	<fct>	<i
1	0500000US01073	1073	Jefferson County, Alabama	1	Total population	660
2	0500000US04013	4013	Maricopa County, Arizona	1	Total population	4087
3	0500000US04019	4019	Pima County, Arizona	1	Total population	1004
4	0500000US06001	6001	Alameda County, California	1	Total population	1610
5	0500000US06013	6013	Contra Costa County, California	1	Total population	1111
6	0500000US06019	6019	Fresno County, California	1	Total population	965

Task 1: Data fields, their data type, and intent:

Id: Data Type: varchar (contains text and numbers) Intent: unique identifier for each row

Id2: Data Type: integer Intent: identifier for geography

Geography: Data Type: varchar (contains text) Intent: geographical location

PopGroupID: Data Type: integer Intent: identifier for population groups

POPGROUP.display-label: Data Type: varchar (contains text) Intent: display label for population groups

RacesReported: Data Type: integer Intent: total races reported

HSDegree: Data Type: float Intent: percentage of the population with a high school degree

BachDegree: Data Type: float Intent: percentage of the population with a bachelor's degree

Task 2:Functions and their results

In [16]:
```
# Load necessary library
library(ggplot2)
```

In [17]:
```
# Display structure of the data frame
str(df)
```

```
'data.frame':   136 obs. of  1 variable:
 $ HSDegree: num  89.1 86.8 88 86.9 88.8 73.6 74.5 77.5 84.6 80.6 ...
```

In [9]:
```
# Display number of rows
nrow(df)
```

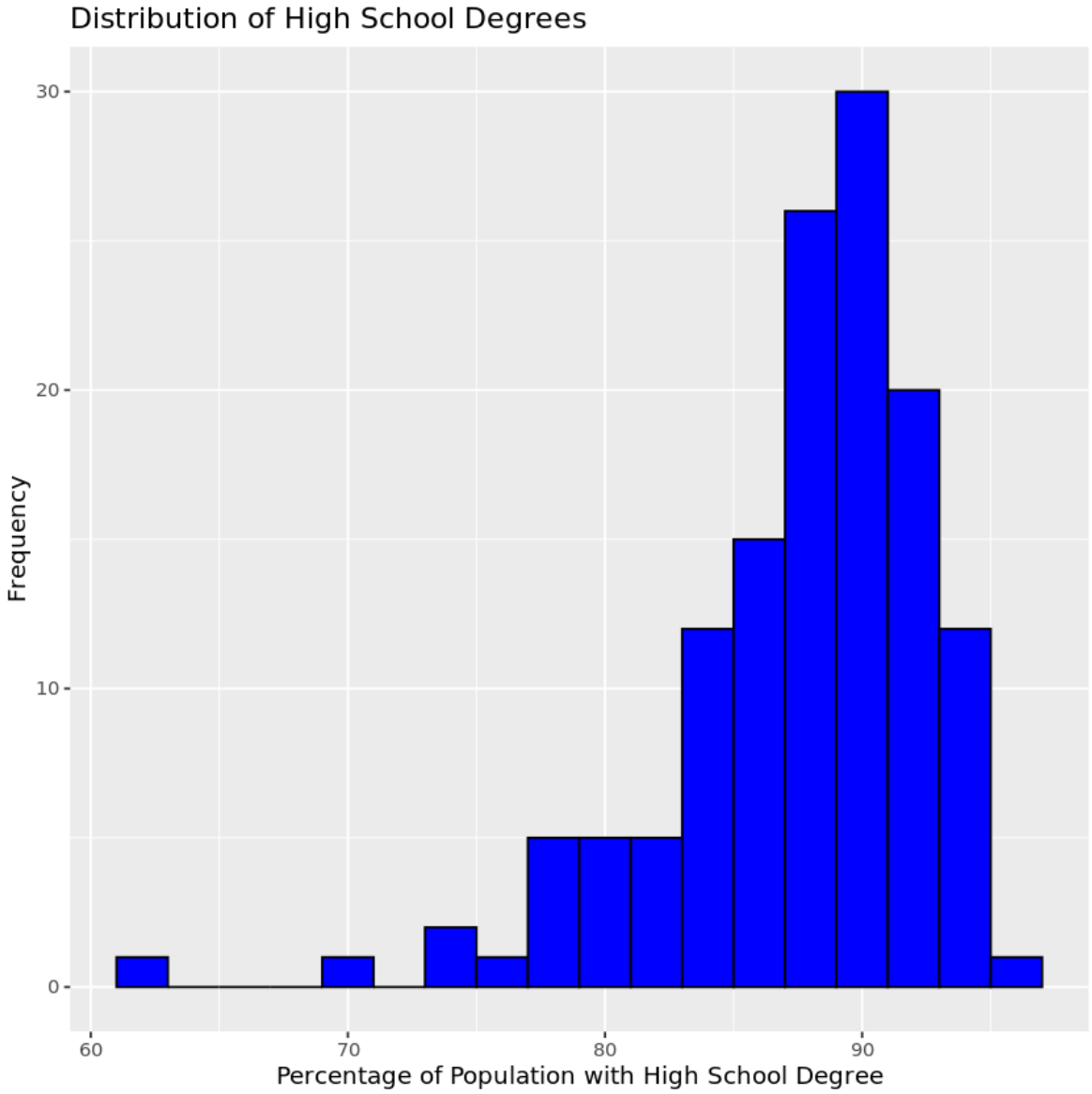
136

In [10]:
```
# Display number of columns
ncol(df)
```

8

Task 3: Histogram distribution of the HSDegree variable

In [11]:
```
## Create Histogram
ggplot(df, aes(x = HSDegree)) +
  geom_histogram(binwidth = 2, fill = "blue", color = "black") +
  labs(title = "Distribution of High School Degrees",
       x = "Percentage of Population with High School Degree",
       y = "Frequency")
```



In [29]:
```r
# Summary Statistics and Skewness

# Install the moments package
if (!requireNamespace("moments", quietly = TRUE)) {
  install.packages("moments")
}

# Load the moments package
library(moments)

# Create the dataframe
df <- data.frame(HSDegree = c(89.1, 86.8, 88, 86.9, 88.8, 73.6, 74.5, 77.5, 84.6

# Summary statistics
summary_stats <- summary(df$HSDegree)
print(summary_stats)

# Check skewness using moments package
skewness_value <- skewness(df$HSDegree)
print(paste("Skewness:", skewness_value))
```

```
Updating HTML index of packages in '.Library'
Making 'packages.html' ... done
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  62.20   85.50   88.70   87.63   90.75   95.50 
[1] "Skewness: -1.69340954614526"
```

Task 4: Answers based on the Histogram

4.1. Is the data distribution unimodal?

Yes, the data distribution is unimodal because there is one clear peak in the histogram.

4.2. Is it approximately symmetrical?

No, the skewness value of -1.69 indicates a left-skewed (negatively skewed) distribution. This means that the left tail is longer or fatter than the right tail.

4.3. Is it approximately bell-shaped?

The histogram may suggest that it is not perfectly bell-shaped, as it is skewed.

4.4. Is it approximately normal?

No, the skewness value being significantly away from 0 and the visual inspection of the histogram suggest that the distribution is not normal.

4.5. If not normal, is the distribution skewed?

Yes, the distribution is left-skewed, as indicated by the negative skewness value (-1.69).

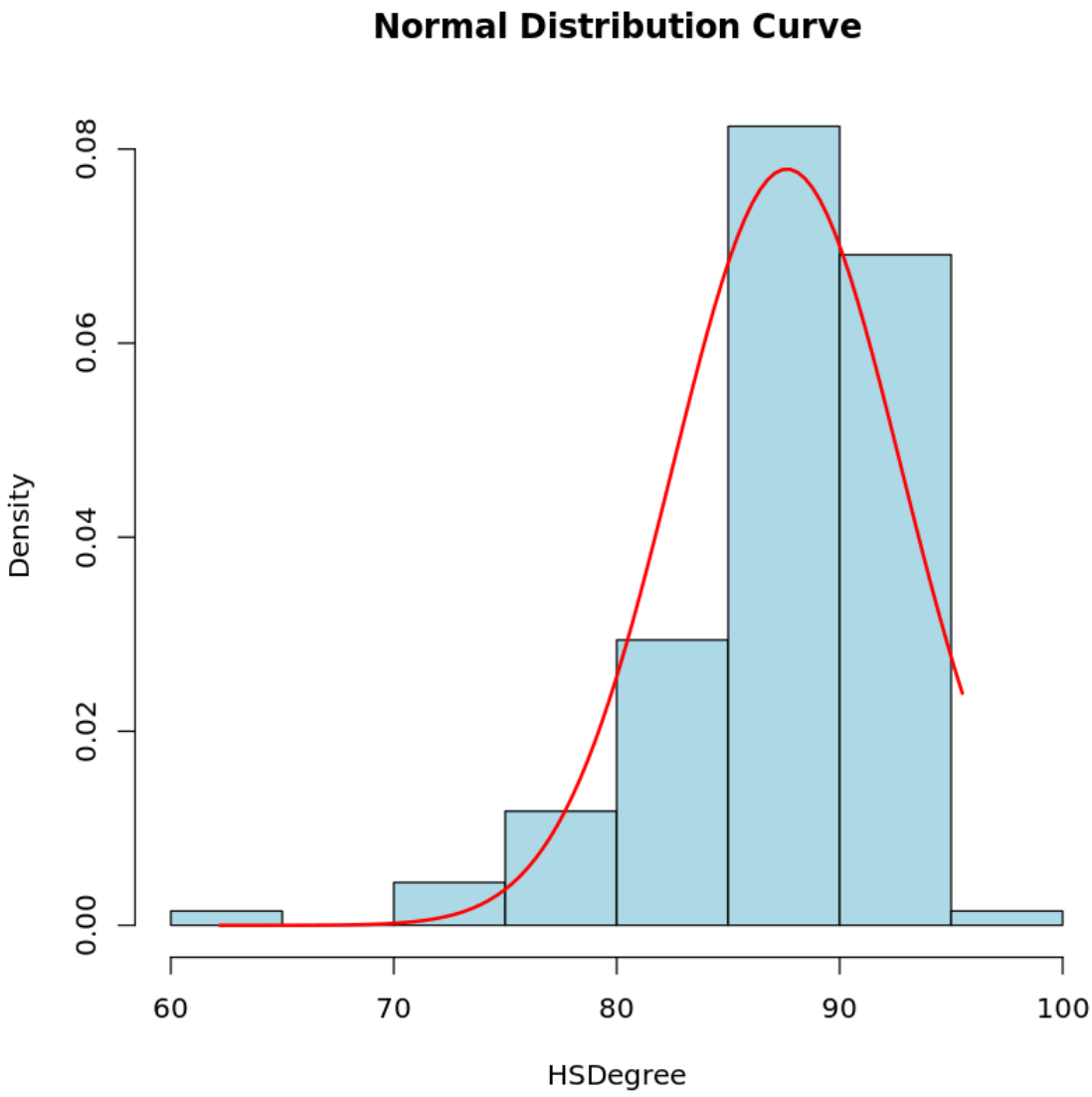
Task 5: Normal curve and Histogram

In [20]:
```r
# 5.1. A normal curve to the Histogram.

# the dataset
HSDegree <- c(89.1, 86.8, 88, 86.9, 88.8, 73.6, 74.5, 77.5, 84.6, 80.6, 86.8, 78

# Plot histogram
hist(HSDegree, freq = FALSE, col = "lightblue", main = "Normal Distribution Curv

# Add normal distribution curve
mu <- mean(HSDegree)
sigma <- sd(HSDegree)
x <- seq(min(HSDegree), max(HSDegree), length = 100)
y <- dnorm(x, mean = mu, sd = sigma)
lines(x, y, col = "red", lwd = 2)
```



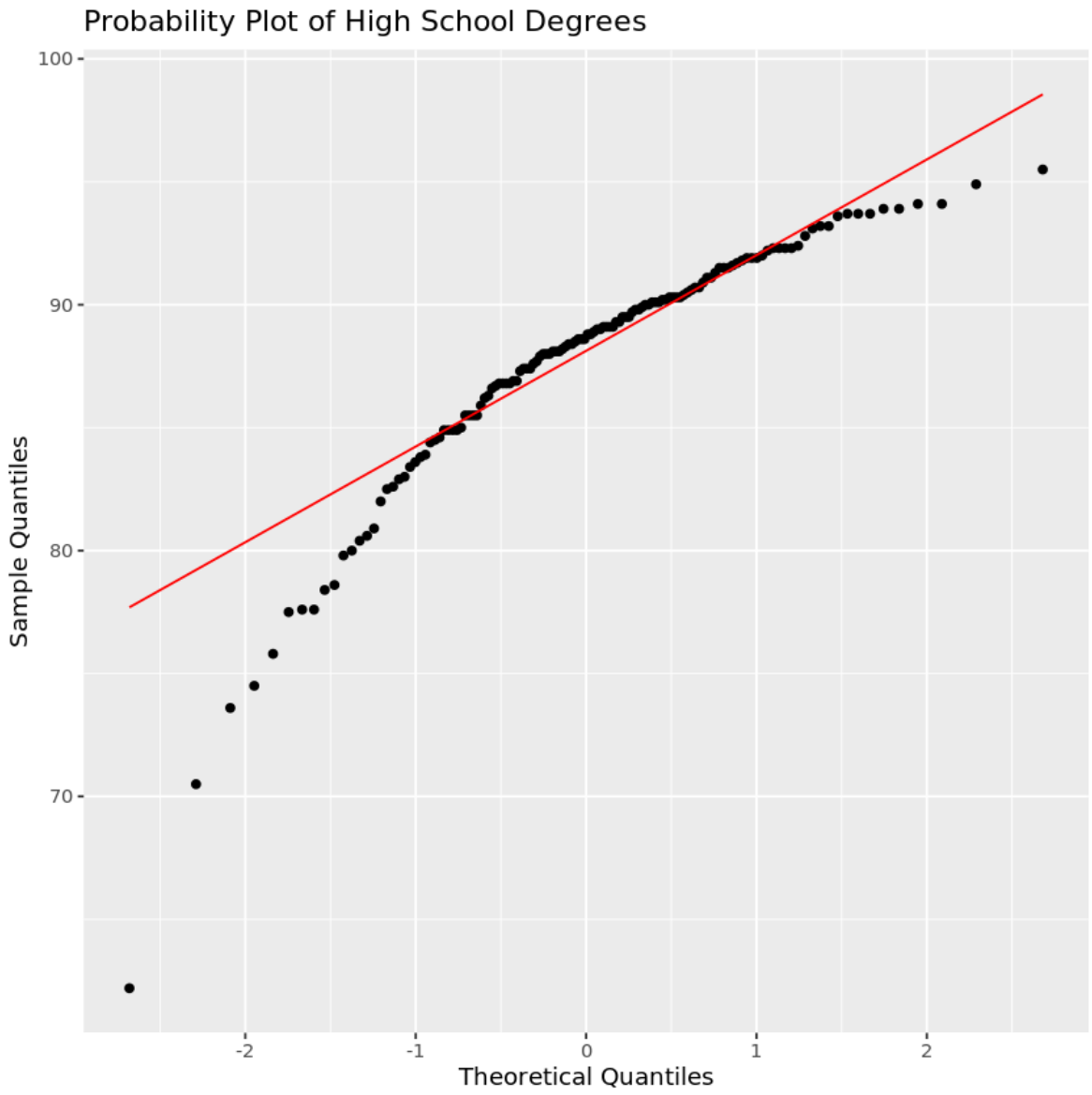
5.2. Explain whether a normal distribution can accurately be used as a model for this data.

Given the left-skewed nature of the data, it deviates from a normal distribution. Therefore, a normal distribution may not accurately represent this data. Instead, a distribution with a left-skewed shape might be a more appropriate model.

Task 6: Probability Plot of the HSDegree variable

In [21]:
```
# Load necessary library
library(ggplot2)

ggplot(df, aes(sample = HSDegree)) +
  geom_qq() +
  geom_qq_line(color = "red") +
  labs(title = "Probability Plot of High School Degrees",
       x = "Theoretical Quantiles",
       y = "Sample Quantiles")
```



Task 7: Descriptive statistics

In [32]:
```
# Install and load the pastecs package
install.packages("pastecs")
library(pastecs)
```

```
Updating HTML index of packages in '.Library'
Making 'packages.html' ... done
```

In [33]:
```
# Display descriptive statistics
stat.desc(df$HSDegree)
```

nbr.val: 136 **nbr.null:** 0 **nbr.na:** 0 **min:** 62.2 **max:** 95.5 **range:** 33.3 **sum:** 11918 **median:** 88.7 **mean:** 87.6323529411765 **SE.mean:** 0.438859785193231 **CI.mean.0.95:** 0.867929607967526 **var:** 26.1933159041394 **std.dev:** 5.11794059208774 **coef.var:** 0.0584024098442636

Task 8: Explanation of skew, kurtosis, and z-scores.

Skewness: Indicates the asymmetry of the data distribution. A skewness value near zero suggests symmetry.

Kurtosis: Measures the tailedness of the data distribution. Positive kurtosis indicates heavier tails.

Z-scores: Measure how many standard deviations a data point is from the mean.

Task 9: Explain how a change in the sample size may change your explanation.

Skew and Kurtosis: These are sample statistics and are influenced by sample size. Larger samples tend to stabilize these measures.

Z-scores: Generally, the z-scores become more reliable with larger sample sizes as the standard error decreases.