```
> # Assignment: ASSIGNMENT 3.2
> # Name: Koppula, Veera
> # Date: 2010-06-21
>
> ## Load the ggplot2 package
> library(ggplot2)
> theme_set(theme_minimal())
```

1. For this exercise, you will use the following dataset, [2014 American Community Survey](). This data is maintained by the US Census Bureau and are designed to show how communities are changing. Through asking questions of a sample of the population, it produces national data on more than 35 categories of information, such as education, income, housing, and employment. For this assignment, you will need to load and activate the ggplot2 package. For this deliverable, you should provide the following:
    1. What are the elements in your data (including the categories and data types)?

Total Elements in the data:8
Id – String element variable characters(13 Characters)
Id2 - Number(4 Digits)
Geography – County name, State name
PopGroupID – a digit
POPGROUP.display.label – character label "Total Population"
Races Reports – Count number
HSDegree – Percentage of High School Degree graduates
BachDegree – Percentage of Bachelor Degree graduates

```
> acs_df <- read.csv("data/acs-14-1yr-s0201.csv")
> acs_df
```

	Id	Id2	Geography	PopGroupID	POPGROUP.display.label	RacesReported	HSDegree	BachDegree
1	0500000US01073	1073	Jefferson County, Alabama	1	Total population	660793	89.1	30.5
2	0500000US04013	4013	Maricopa County, Arizona	1	Total population	4087191	86.8	30.2
3	0500000US04019	4019	Pima County, Arizona	1	Total population	1004516	88.0	30.8
4	0500000US06001	6001	Alameda County, California	1	Total population	1610921	86.9	42.8
5	0500000US06013	6013	Contra Costa County, California	1	Total population	1111339	88.8	39.7
6	0500000US06019	6019	Fresno County, California	1	Total population	965974	73.6	19.7
7	0500000US06029	6029	Kern County, California	1	Total population	874589	74.5	15.4
8	0500000US06037	6037	Los Angeles County, California	1	Total population	10116705	77.5	30.3
9	0500000US06059	6059	Orange County, California	1	Total population	3145515	84.6	38.0
10	0500000US06065	6065	Riverside County, California	1	Total population	2329271	80.6	20.7
11	0500000US06067	6067	Sacramento County, California	1	Total population	1482026	86.8	28.9
12	0500000US06071	6071	San Bernardino County, California	1	Total population	2112619	78.6	18.9
13	0500000US06073	6073	San Diego County, California	1	Total population	3263431	86.6	37.1
14	0500000US06075	6075	San Francisco County, California	1	Total population	852469	88.1	54.2
15	0500000US06077	6077	San Joaquin County, California	1	Total population	715597	77.6	18.3
16	0500000US06081	6081	San Mateo County, California	1	Total population	758581	88.1	47.5
17	0500000US06085	6085	Santa Clara County, California	1	Total population	1894605	87.4	48.4
18	0500000US06097	6097	Sonoma County, California	1	Total population	500292	87.6	34.8
19	0500000US06099	6099	Stanislaus County, California	1	Total population	531997	78.4	17.0
20	0500000US06111	6111	Ventura County, California	1	Total population	846178	83.6	31.6
21	0500000US08005	8005	Arapahoe County, Colorado	1	Total population	618821	91.9	40.9
22	0500000US08031	8031	Denver County, Colorado	1	Total population	663862	85.5	44.3
23	0500000US08041	8041	El Paso County, Colorado	1	Total population	663519	92.8	36.5
24	0500000US08059	8059	Jefferson County, Colorado	1	Total population	558503	94.1	42.0
25	0500000US09001	9001	Fairfield County, Connecticut	1	Total population	945438	89.8	46.7

26	0500000US09003	9003	Hartford County, Connecticut	1	Total population	897985	89.3	36.8	
27	0500000US09009	9009	New Haven County, Connecticut	1	Total population	861277	89.5	34.5	
28	0500000US10003	10003	New Castle County, Delaware	1	Total population	552778	90.1	35.8	
29	0500000US11001	11001	District of Columbia, District of Columbia	1	Total population	658893	90.2	55.0	
30	0500000US12009	12009	Brevard County, Florida	1	Total population	556885	91.6	27.2	
31	0500000US12011	12011	Broward County, Florida	1	Total population	1869235	88.4	30.5	
32	0500000US12031	12031	Duval County, Florida	1	Total population	897698	89.0	26.1	
33	0500000US12057	12057	Hillsborough County, Florida	1	Total population	1316298	87.3	29.8	
34	0500000US12071	12071	Lee County, Florida	1	Total population	679513	86.3	26.5	
35	0500000US12086	12086	Miami-Dade County, Florida	1	Total population	2662874	80.9	26.6	
36	0500000US12095	12095	Orange County, Florida	1	Total population	1253001	87.9	31.4	
37	0500000US12099	12099	Palm Beach County, Florida	1	Total population	1397710	87.7	33.0	
38	0500000US12103	12103	Pinellas County, Florida	1	Total population	938098	90.1	29.5	
39	0500000US12105	12105	Polk County, Florida	1	Total population	634638	84.9	19.7	
40	0500000US12127	12127	Volusia County, Florida	1	Total population	507531	88.9	22.5	
41	0500000US13067	13067	Cobb County, Georgia	1	Total population	730981	90.3	43.7	
42	0500000US13089	13089	DeKalb County, Georgia	1	Total population	722161	88.4	41.7	
43	0500000US13121	13121	Fulton County, Georgia	1	Total population	996319	91.3	49.2	
44	0500000US13135	13135	Gwinnett County, Georgia	1	Total population	877922	88.0	35.4	
45	0500000US15003	15003	Honolulu County, Hawaii	1	Total population	991788	91.8	32.6	
46	0500000US17031	17031	Cook County, Illinois	1	Total population	5246456	85.5	36.2	
47	0500000US17043	17043	DuPage County, Illinois	1	Total population	932708	92.3	48.0	
48	0500000US17089	17089	Kane County, Illinois	1	Total population	527306	82.9	32.6	
49	0500000US17097	17097	Lake County, Illinois	1	Total population	705186	90.3	44.0	
50	0500000US17197	17197	Will County, Illinois	1	Total population	685419	90.7	33.1	
51	0500000US18097	18097	Marion County, Indiana	1	Total population	934243	85.0	28.8	
52	0500000US20091	20091	Johnson County, Kansas	1	Total population	574272	95.5	52.8	
53	0500000US20173	20173	Sedgwick County, Kansas	1	Total population	508803	88.8	30.7	
54	0500000US21111	21111	Jefferson County, Kentucky	1	Total population	760026	88.5	31.6	
55	0500000US24003	24003	Anne Arundel County, Maryland	1	Total population	560133	91.9	38.8	
56	0500000US24005	24005	Baltimore County, Maryland	1	Total population	826925	90.4	37.2	
57	0500000US24031	24031	Montgomery County, Maryland	1	Total population	1030447	90.9	58.5	
58	0500000US24033	24033	Prince George's County, Maryland	1	Total population	904430	85.5	31.0	
59	0500000US24510	24510	Baltimore city, Maryland	1	Total population	622793	84.4	30.0	
60	0500000US25005	25005	Bristol County, Massachusetts	1	Total population	554194	82.5	25.7	
61	0500000US25009	25009	Essex County, Massachusetts	1	Total population	769091	89.1	38.9	
62	0500000US25017	25017	Middlesex County, Massachusetts	1	Total population	1570315	92.3	52.3	
63	0500000US25021	25021	Norfolk County, Massachusetts	1	Total population	692254	94.1	51.9	
64	0500000US25023	25023	Plymouth County, Massachusetts	1	Total population	507022	92.2	34.1	
65	0500000US25025	25025	Suffolk County, Massachusetts	1	Total population	767254	83.9	42.3	
66	0500000US25027	25027	Worcester County, Massachusetts	1	Total population	813475	90.1	34.6	
67	0500000US26081	26081	Kent County, Michigan	1	Total population	629237	89.1	33.7	
68	0500000US26099	26099	Macomb County, Michigan	1	Total population	860112	89.3	23.9	

69	0500000US26125	26125	Oakland County, Michigan	1	Total population	1237868	93.6	44.8
70	0500000US26163	26163	Wayne County, Michigan	1	Total population	1764804	84.9	22.1
71	0500000US27053	27053	Hennepin County, Minnesota	1	Total population	1212064	93.2	47.3
72	0500000US27123	27123	Ramsey County, Minnesota	1	Total population	532655	89.9	40.9
73	0500000US29095	29095	Jackson County, Missouri	1	Total population	683191	90.0	29.5
74	0500000US29189	29189	St. Louis County, Missouri	1	Total population	1001876	93.2	42.8
75	0500000US31055	31055	Douglas County, Nebraska	1	Total population	543244	88.2	36.3
76	0500000US32003	32003	Clark County, Nevada	1	Total population	2069681	84.5	22.7
77	0500000US34003	34003	Bergen County, New Jersey	1	Total population	933572	91.5	46.2
78	0500000US34007	34007	Camden County, New Jersey	1	Total population	511038	88.3	31.3
79	0500000US34013	34013	Essex County, New Jersey	1	Total population	795723	85.5	32.7
80	0500000US34017	34017	Hudson County, New Jersey	1	Total population	669115	83.4	38.2
81	0500000US34023	34023	Middlesex County, New Jersey	1	Total population	836297	89.1	41.0
82	0500000US34025	34025	Monmouth County, New Jersey	1	Total population	629279	93.1	43.7
83	0500000US34029	34029	Ocean County, New Jersey	1	Total population	586301	91.7	28.6
84	0500000US34031	34031	Passaic County, New Jersey	1	Total population	508856	83.8	28.6
85	0500000US34039	34039	Union County, New Jersey	1	Total population	552939	86.2	33.0
86	0500000US35001	35001	Bernalillo County, New Mexico	1	Total population	675551	88.0	32.7
87	0500000US36005	36005	Bronx County, New York	1	Total population	1438159	70.5	19.3
88	0500000US36029	36029	Erie County, New York	1	Total population	922835	90.6	31.3
89	0500000US36047	36047	Kings County, New York	1	Total population	2621793	80.0	34.3
90	0500000US36055	36055	Monroe County, New York	1	Total population	749857	90.3	35.9
91	0500000US36059	36059	Nassau County, New York	1	Total population	1358627	90.7	43.2
92	0500000US36061	36061	New York County, New York	1	Total population	1636268	86.8	59.9
93	0500000US36081	36081	Queens County, New York	1	Total population	2321580	80.4	29.8
94	0500000US36103	36103	Suffolk County, New York	1	Total population	1502968	89.8	34.0
95	0500000US36119	36119	Westchester County, New York	1	Total population	972634	87.4	47.1
96	0500000US37081	37081	Guilford County, North Carolina	1	Total population	512119	89.0	33.3
97	0500000US37119	37119	Mecklenburg County, North Carolina	1	Total population	1012539	89.5	43.0
98	0500000US37183	37183	Wake County, North Carolina	1	Total population	998691	92.4	49.2
99	0500000US39035	39035	Cuyahoga County, Ohio	1	Total population	1259828	88.1	31.0
100	0500000US39049	39049	Franklin County, Ohio	1	Total population	1231393	90.0	38.0
101	0500000US39061	39061	Hamilton County, Ohio	1	Total population	806631	90.5	35.6
102	0500000US39113	39113	Montgomery County, Ohio	1	Total population	533116	89.7	25.7
103	0500000US39153	39153	Summit County, Ohio	1	Total population	541943	91.1	30.3
104	0500000US40109	40109	Oklahoma County, Oklahoma	1	Total population	766215	86.8	30.6
105	0500000US40143	40143	Tulsa County, Oklahoma	1	Total population	629598	88.6	30.7

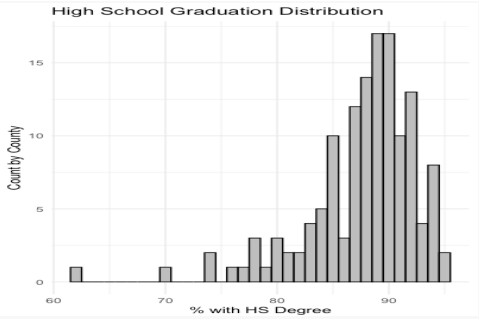
```
106 0500000US41051 41051          Multnomah County, Oregon       1      Total population       776712     91.1
41.6
107 0500000US41067 41067          Washington County, Oregon      1      Total population       562998     90.2
39.7
108 0500000US42003 42003          Allegheny County, Pennsylvania 1      Total population       1231255    93.9
37.7
109 0500000US42017 42017          Bucks County, Pennsylvania     1      Total population       626685     93.9
37.7
110 0500000US42029 42029          Chester County, Pennsylvania   1      Total population       512784     92.3
49.3
111 0500000US42045 42045          Delaware County, Pennsylvania  1      Total population       562960     91.5
36.3
112 0500000US42071 42071          Lancaster County, Pennsylvania 1      Total population       533320     84.9
26.0
113 0500000US42091 42091          Montgomery County, Pennsylvania 1     Total population       816857     93.7
47.3
114 0500000US42101 42101          Philadelphia County, Pennsylvania 1   Total population       1560297    82.6
26.0
115 0500000US44007 44007          Providence County, Rhode Island 1     Total population       631974     82.0
25.2
116 0500000US47037 47037          Davidson County, Tennessee     1      Total population       668347     86.7
37.3
117 0500000US47157 47157          Shelby County, Tennessee       1      Total population       938803     87.4
29.9
118 0500000US48029 48029          Bexar County, Texas            1      Total population       1855866    83.0       26.3
119 0500000US48085 48085          Collin County, Texas           1      Total population       885241     93.7       50.0
120 0500000US48113 48113          Dallas County, Texas           1      Total population       2518638    77.6       29.1
121 0500000US48121 48121          Denton County, Texas           1      Total population       753363     91.9       41.5
122 0500000US48141 48141          El Paso County, Texas          1      Total population       833487     75.8       21.1
123 0500000US48157 48157          Fort Bend County, Texas        1      Total population       685345     88.6       44.1
124 0500000US48201 48201          Harris County, Texas           1      Total population       4441370    79.8       29.7
125 0500000US48215 48215          Hidalgo County, Texas          1      Total population       831073     62.2       17.9
 [ reached 'max' / getOption("max.print") -- omitted 11 rows ]
>
> class(acs_df)
[1] "data.frame"
```

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

```
> ##str() result
        str(acs_df)
        '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() result
        > nrow(acs_df)
        [1] 136
> ##ncol() result
        > ncol(acs_df)
        [1] 8
```

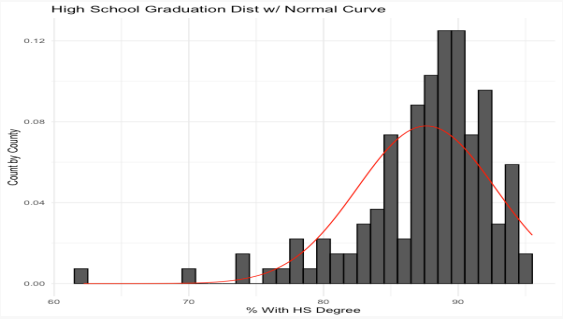
3. Create a Histogram of the HSDegree variable using the ggplot2 package.
    1. Set a bin size for the Histogram.
    2. Include a Title and appropriate X/Y axis labels on your Histogram Plot.

```
> ggplot(acs_df, aes(HSDegree)) + geom_histogram(colour = "black", fill = "grey", binwidth = 1,bins = 1) +
ggtitle("High School Graduation Distribution") + labs(x = "% with HS Degree", y = "Count by County")
```



4. Answer the following questions based on the Histogram produced:
    1. Based on what you see in this histogram, is the data distribution unimodal?
       Yes, the distribution is Unimodal
    2. Is it approximately symmetrical?
       No, the distribution is not symmetrical, distribution is skewed left (negative)
    3. Is it approximately bell-shaped?
       Yes, the distribution is approximately bell-shaped, but with negative Skewed
    4. Is it approximately normal?
       No, this distribution is not normal, as it has negative skew (skewed left)
    5. If not normal, is the distribution skewed? If so, in which direction?
       Distribution is skewed left (negative skew)
    6. Include a normal curve to the Histogram that you plotted.

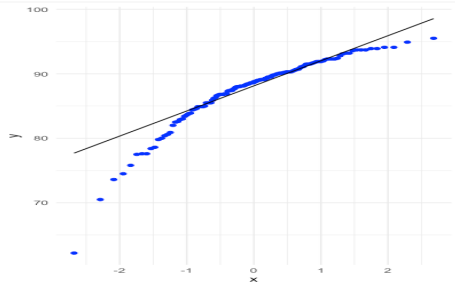
```
> ggplot(data = acs_df) + geom_histogram(mapping = aes(x = HSDegree, y=..density..),color="black", binwidth = 1,bins = 1) +
ggtitle("High School Graduation Dist w/ Normal Curve") + xlab("% With HS Degree") + ylab("Count by County")+
stat_function(fun = dnorm, color = "red", args = list(mean = mean(acs_df$HSDegree), sd = sd(acs_df$HSDegree)))
```



    7. Explain whether a normal distribution can accurately be used as a model for this data.
       No, Normal distribution can not be used as model for this data, as the data is heavily negative/Left skewed

5. Create a Probability Plot of the HSDegree variable.

```
> ggplot(acs_df, aes(sample=HSDegree)) + stat_qq(col="blue") + stat_qq_line(col="black")
```



6. 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.

       Distribution is not normal, this can be validated as the data plot is deviating from straight line
    2. If not normal, is the distribution skewed? If so, in which direction? Explain how you know.

       This distribution is left skewed. Its left skewed as plotted points bend down and to the right of the normal line that indicates a long tail to the left.
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.

```
> stat.desc(acs_df$HSDegree, basic=FALSE, norm=TRUE)
       median          mean       SE.mean  CI.mean.0.95           var       std.dev      coef.var      skewness      skew.2SE      kurtosis      kurt.2SE  normtest.W
8.870000e+01  8.763235e+01  4.388598e-01  8.679296e-01  2.619332e+01  5.117941e+00  5.840241e-02 -1.674767e+00 -4.030254e+00  4.352856e+00  5.273885e+00  8.773635e-01
   normtest.p
3.193634e-09
>
> stat.desc(acs_df$BachDegree, basic = FALSE, norm = TRUE)
     median          mean       SE.mean CI.mean.0.95           var       std.dev      coef.var      skewness      skew.2SE      kurtosis
kurt.2SE   normtest.W   normtest.p
 34.10000000  35.46102941   0.81545273   1.61271456  90.43498856   9.50973126   0.26817415   0.32843046   0.79035382
-0.27742492  -0.33612576   0.98316075   0.09206162
```

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 for HS Degree is negative, indicating the values piled up on the right of distribution
   Kurtosis for HS Degree is positive, indicating the values are pointy and heavy- tailed distribution
   The value of skew.2SE and kurt.2SE are equal to skew and kurtosis divided by 2 standard errors. By normalizing skew and kurtosis in this way. In current case skew.2SE < 0 and kurtosis.2SE > 0 indicating that data is distorted, and data is skewed to the right.
   Because these normalized values involve dividing by 2 standard errors, they are sensitive to the size of the sample. skew.2SE and kurt.2SE are most appropriate for relatively small samples, 30-50. For larger samples, it is best to compute values corresponding to 2.58SE ($p < 0.01$) and 3.29SE ($p < 0.001$). In very large samples, say 200 observations or more, it is best to look at the shape of the distribution visually and consider the actual values of skew and kurtosis, not their normalized values.