Justin Madsen

DSC 520

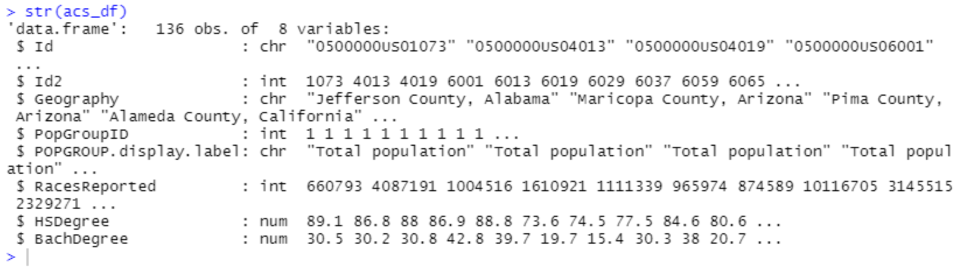
Dr. Bushart

Assignmen03

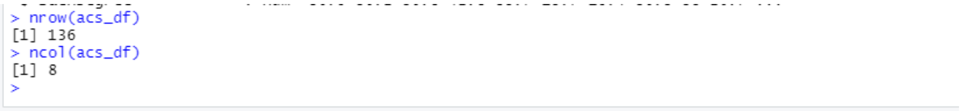
A

i - Id(alphanumeric string), Id2 (numeric, int), Geography (string), PopGroupID (numeric, int), POPGROUP.display.label (string), RacesReported(numeric), HSdegree(numeric, float), BachDegree(numeric, float)

ii -

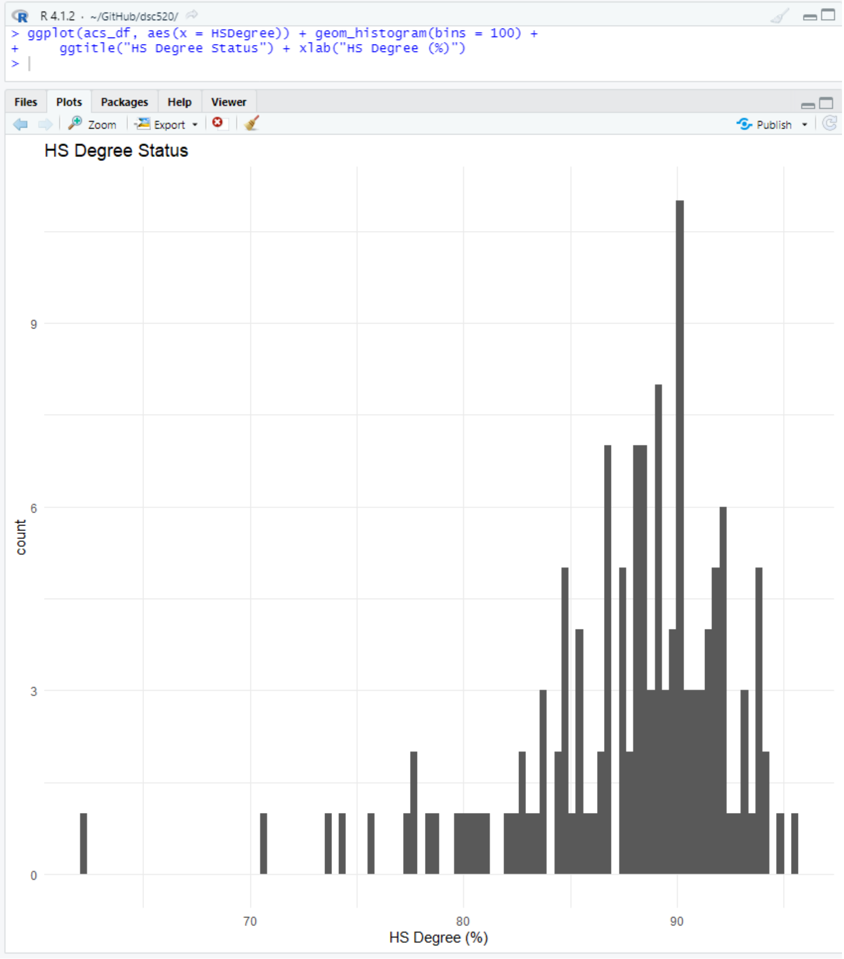


```
> str(acs_df)
'data.frame':	136 obs. of  8 variables:
 $ Id                    : chr  "0500000US01073" "0500000US04013" "0500000US04019" "0500000US06001"
 ...
 $ Id2                   : int  1073 4013 4019 6001 6013 6019 6029 6037 6059 6065 ...
 $ Geography             : chr  "Jefferson County, Alabama" "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 popul
ation" ...
 $ 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(acs_df)
[1] 136
> ncol(acs_df)
[1] 8
>
```

iii -

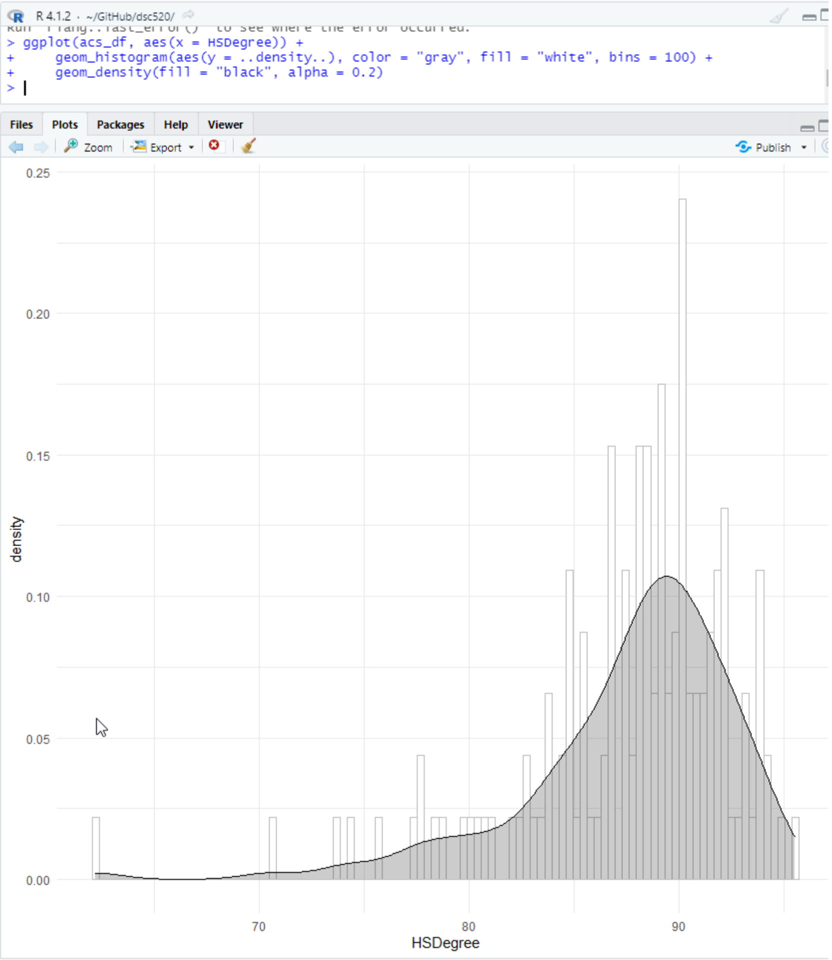


```
> ggplot(acs_df, aes(x = HSDegree)) + geom_histogram(bins = 100) +
+     ggtitle("HS Degree Status") + xlab("HS Degree (%)")
> |
```

iv –

1. This data is unimodal.
2. The data is not symmetrical.
3. The data is bell shaped.
4. This data is not normal.
5. This data is skewed negatively.

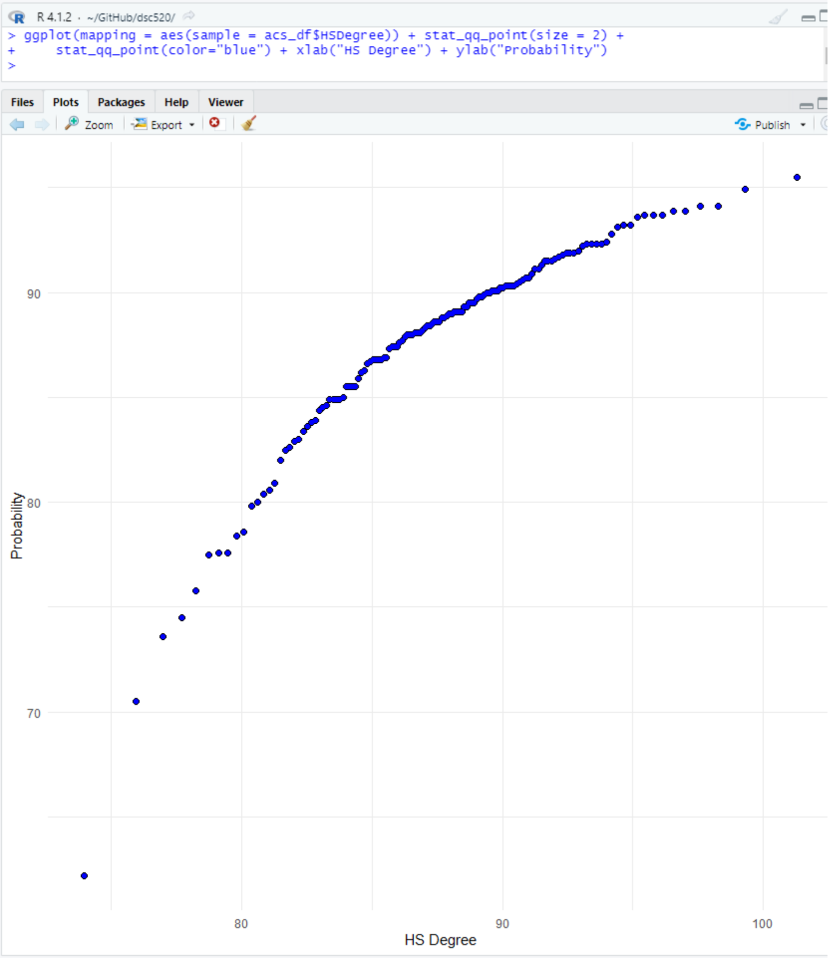
6.



```
Run `rlang::last_error()` to see where the error occurred.
> ggplot(acs_df, aes(x = HSDegree)) +
+     geom_histogram(aes(y = ..density..), color = "gray", fill = "white", bins = 100) +
+     geom_density(fill = "black", alpha = 0.2)
> 
```

7. A normal distribution can’t be used, the outliers drag the curve down while also skewing the upper end of the data to be lower than expected.

v -

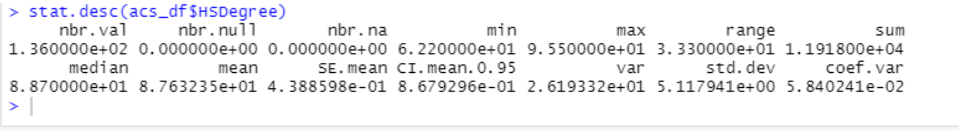


```
R 4.1.2 · ~/GitHub/dsc520/
> ggplot(mapping = aes(sample = acs_df$HSDegree)) + stat_qq_point(size = 2) +
+     stat_qq_point(color="blue") + xlab("HS Degree") + ylab("Probability")
>
```

vi -

1. The plot graph does not appear to be normal. If it were normal, I would expect a straighter line.
2. This data looks skewed to the left. Using https://www.quality-control-plan.com/StatGuide/probplots.htm as a reference, I noticed that there are outliers on the bottom end. I don't believe the top end are outliers, however depending on the person they might be. Since the data bows up above a straight line, this would be skewing to the left.

vii -



```
> stat.desc(acs_df$HSDegree)
     nbr.val     nbr.null       nbr.na          min          max        range          sum
1.360000e+02 0.000000e+00 0.000000e+00 6.220000e+01 9.550000e+01 3.330000e+01 1.191800e+04
      median         mean      SE.mean CI.mean.0.95          var      std.dev     coef.var
8.870000e+01 8.763235e+01 4.388598e-01 8.679296e-01 2.619332e+01 5.117941e+00 5.840241e-02
>
```

viii – The skewness of the data is negative skewed, while the probability is skewed to the left. The kurtosis of this data would be leptokurtic, due to data falling outside of the 3 standard deviations. Using the moments package, skewness and kurtosis functions support this theory with a skewness of –1.69 and a kurtosis of 7.46. Z-score is defined as how many standard deviations from the mean a data point is. Due to the abnormal nature of this data set, many of the outliers will fall closer to the -6 or -7 z-score range. The large grouping around 90 will also fall between –1 and 1. As data is added, outliers may become more normalized or even further pushed to the edges. This would depend on the data being ingested. If the data is more normalizing, then the kurtosis would be expected to decrease as there are fewer datapoints outside of the standard. The skews would also slowly approach normal.