**Topic – 3**

AUTOMOBILES

Name :- Akash Mandal

Roll No. :- 16MA20007

Course :- CS40003

Problem Statement –

a) Categorize all the attributes listed in the table according to the NOIR topology?

b) Apply the applicable central tendency measures to any four attributes taking one attribute from each category.

c) Consider the attribute “peak-rpm” and “city-mpg”? Find which probability distribution(s) they are likely to follow?

Reference –

AUTOMOBILES.csv (Included in the project folder for reference)

Observation of data –

The database AUTOMOBILES.csv reveals the following basic structure-

1. 205 observations of 26 variables.
2. It is a database about various cars and their technical properties.

Classification of data (NOIR)–

1. symboling – Ordinal: -2, -1, 0, 1, 2, 3
2. normalized-losses – Interval
3. make – Nominal: alfa-romero, audi, bmw… and 19 others.
4. fuel-type – Nominal: diesel, gas
5. aspiration – Nominal: std, turbo
6. num-of-doors – Ordinal: two, four
7. body-style – Nominal: convertible, hardtop, hatchback, sedan, wagon
8. drive-wheels – Nominal: fwd, 4wd, rwd
9. engine-location – Nominal: front, rear
10. wheel-base – Interval
11. length – Ratio
12. width – Ratio
13. height – Ratio
14. curb-weight - Ratio
15. engine-type – Nominal: dohc, dohv, l, ohc, ohcf, ohcv, rotor
16. num-of-cylinders – Ordinal : two, three, four, five, six, eight, twelve
17. engine-size – Interval
18. fuel-system – Nominal: 1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi
19. bore – Interval
20. stroke – Interval
21. compression ratio – Interval
22. horsepower – Ratio
23. peak-rpm – Interval
24. city-mpg – Ratio
25. highway-mpg – Ratio
26. price – Ratio

Code Snippet

> table(auto$symboling)

-2 -1 0 1 2 3

3 22 67 54 32 27

> table(auto$make)

alfa-romero audi bmw chevrolet dodge honda isuzu jaguar mazda mercedes-benz

3 7 8 3 9 13 4 3 17 8

mercury mitsubishi nissan peugot plymouth porsche renault saab subaru toyota

1 13 18 11 7 5 2 6 12 32

volkswagen volvo

12 11

> table(auto$`fuel-type`)

diesel gas

20 185

> table(auto$aspiration)

std turbo

168 37

Code Snippet

> table(auto$`num-of-doors`)

? four two

2 114 89

> table(auto$`body-style`)

convertible hardtop hatchback sedan wagon

6 8 70 96 25

> table(auto$`drive-wheels`)

4wd fwd rwd

9 120 76

> table(auto$`engine-location`)

front rear

202 3

> table(auto$`engine-type`)

dohc dohcv l ohc ohcf ohcv rotor

12 1 12 148 15 13 4

> table(auto$`num-of-cylinders`)

eight five four six three twelve two

5 11 159 24 1 1 4

> table(auto$`fuel-system`)

1bbl 2bbl 4bbl idi mfi mpfi spdi spfi

11 66 3 20 1 94 9 1

Measures of central Tendency

1. Nominal –

We choose the variable ‘body-style’ as an example of Nominal variable. It can take one of five values, namely, ‘convertible’, ‘hardtop’, ‘hatchback’, ‘sedan’ and ‘wagon’.

The most applicable measure of central tendency is the mode, since

1. There is no ordering in the data
2. The data is not numeric

Thus the mode of the data is ‘sedan’, with a total of 96 observations.

Code Snippet

> table(auto$`body-style`)

convertible hardtop hatchback sedan wagon

6 8 70 96 25

1. Ordinal –

We choose the variable num-of-cylinders as our example of an ordinal variable. It can take the values ‘two’, ‘three’, ‘four’, ‘five’, ‘six’, eight’, ‘twelve’

Here too, mode is the best measure for central tendency.

Code Snippet

> table(auto$`num-of-cylinders`)

eight five four six three twelve two

5 11 159 24 1 1 4

The mode of the variable is ‘four’ with a total of 159 observations.

1. Interval -

We choose the variable peak-rpm as an example of Interval type. Mean can be taken as a measure of central tendency. This is because there are no significant outliers which can affect the mean.

The variable is first converted to numeric and the mean is calculated to be – 5125.369

Code Snippet

> auto$`peak-rpm` <- as.numeric(auto$`peak-rpm`)

Warning message:

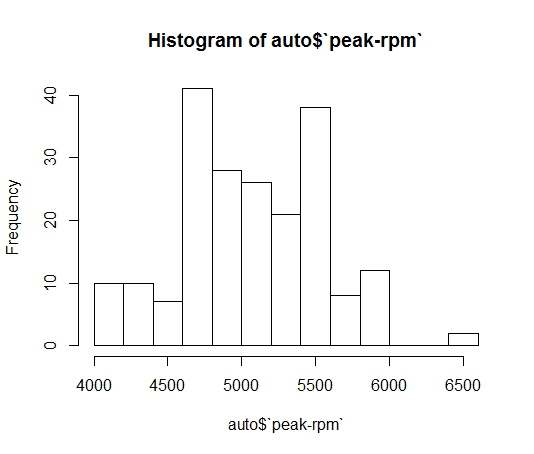
NAs introduced by coercion

> fivenum(auto$`peak-rpm`)

[1] 4150 4800 5200 5500 6600

> mean(auto$`peak-rpm`, na.rm = TRUE)

[1] 5125.369



1. Ratio –

We choose the variable horsepower as an example for Ratio variable. An investigation of the distribution shows a positive skew. Thus the median is preferred over the mean. The mode can also be used but is not used in common.

The median for the data turns out to be – 95hp

Code Snippet

> auto$horsepower <- as.numeric(auto$horsepower)

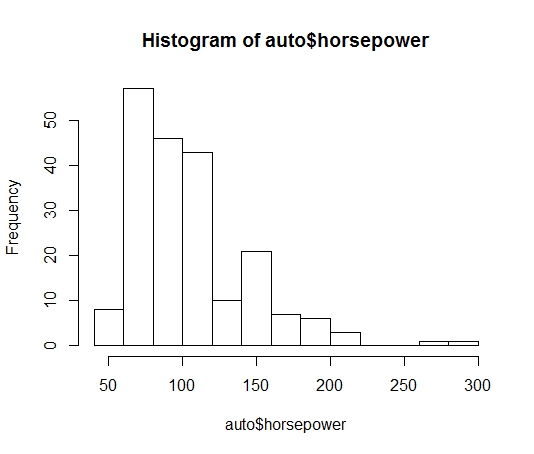
Warning message:

NAs introduced by coercion

> fivenum(auto$horsepower, na.rm = TRUE)

[1] 48 70 95 116 288

> hist(auto$horsepower)



Fitting to distribution –

1. Empirical density, and Cumulative Frequency Distribution –

We use plotdist function of the fitdistrplus package to visualise the density of our data variables, peak-rpm and city-mpg.

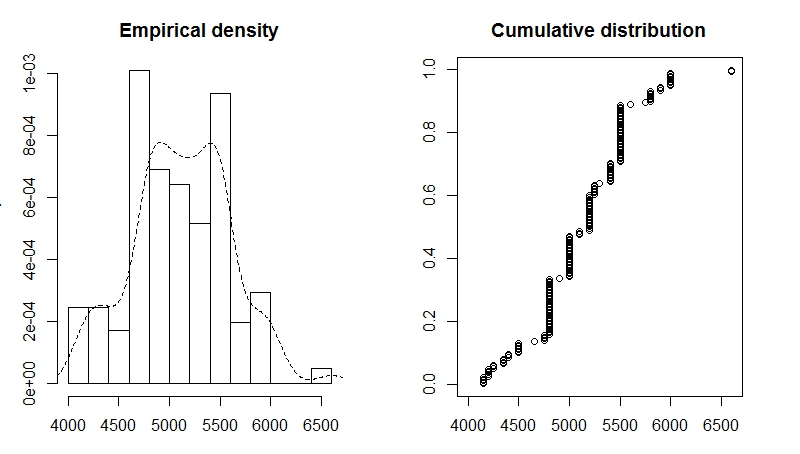
1. Peak-rpm –

The distribution looks approximately symmetric, and has two peaks.

Code Snippet

> rpm <- as.numeric(na.omit(as.numeric(auto$`peak-rpm`)))

> plotdist(rpm, histo = TRUE, demp = TRUE)



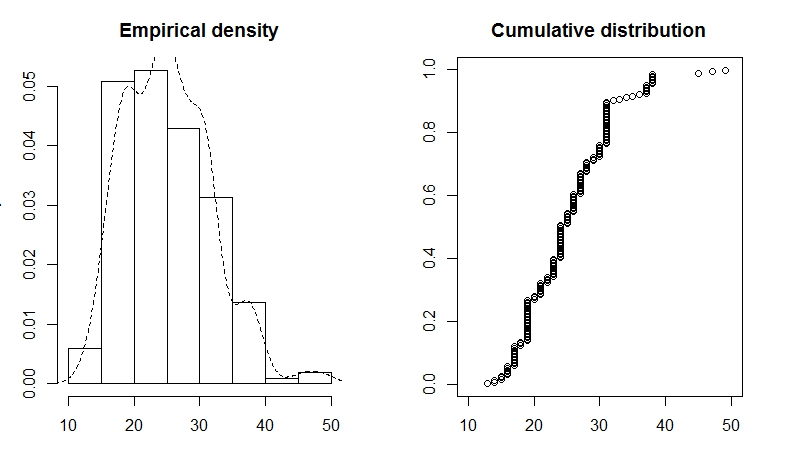
1. City-mpg-

The distribution looks positively skewed.

Code Snippet

> rpm <- as.numeric(na.omit(as.numeric(auto$`peak-rpm`)))

> plotdist(rpm, histo = TRUE, demp = TRUE)



1. Describing the distribution
2. Peak-rpm –

We use descdist function to obtain summary statistics and a Cullen Frey graph.

We see gamma, lognormal, weibull, and normal are probable candidates.

Code Snippet

> descdist(rpm, boot = 500)

summary statistics

------

min: 4150 max: 6600

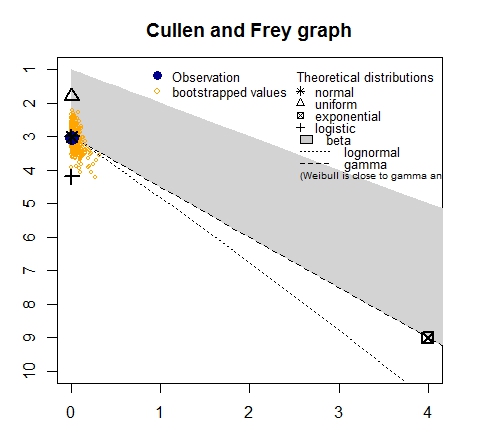
median: 5200

mean: 5125.369

estimated sd: 479.3346

estimated skewness: 0.07323669

estimated kurtosis: 3.056526



1. City-mpg – Shows lognormal, gamma and weibull as probable candidates.

Code Snippet

> descdist(mpg, boot = 500)

summary statistics

------

min: 13 max: 49

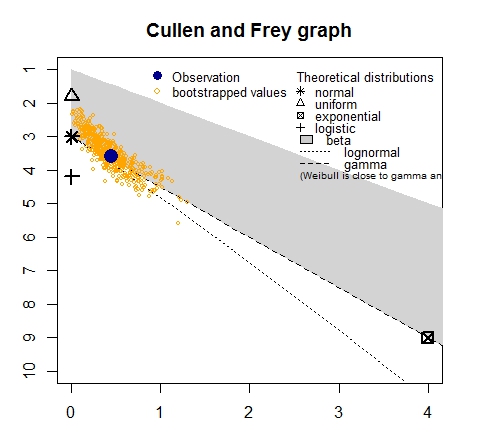
median: 24

mean: 25.21951

estimated sd: 6.542142

estimated skewness: 0.663704

estimated kurtosis: 3.578648



1. Fitting
2. Peak-rpm

After running the fitdist, and the density plot, we find that the normal, gamma, and lognormal give a close fit to the data. The analysis(of other plots – q-q, cdf, p-p) is beyond the scope of this report, and visually, the Normal distribution is a good fit.

Code Snippet

> #fitting to distributions

> rpm\_n <- fitdist(rpm, "norm")

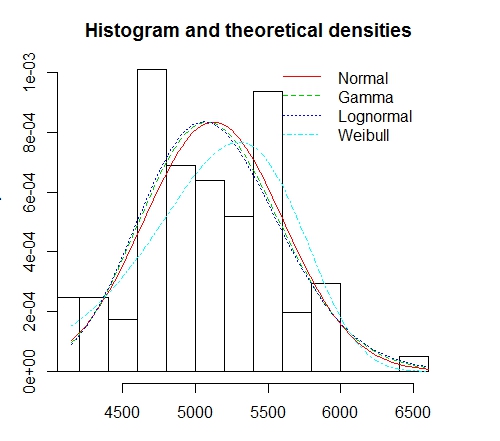
> rpm\_g <- fitdist(rpm, "gamma")

> rpm\_ln <- fitdist(rpm, "lnorm")

> rpm\_w <- fitdist(rpm, "weibull")

> rpm\_plot.legend <- c("Normal", "Gamma", "Lognormal", "Weibull")

> denscomp(list(rpm\_n, rpm\_g, rpm\_ln, rpm\_w), legendtext = rpm\_plot.legend)



1. City-mpg

After running the fitdist, and the density plot, we find that the gamma, and lognormal give a close fit to the data. The analysis(of other plots – q-q, cdf, p-p) is beyond the scope of this report, and visually, the Gamma or Lognormal distributions are good fits.

Code Snippet

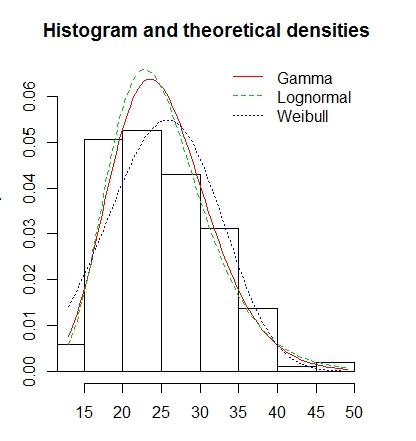
> mpg\_g <- fitdist(mpg, "gamma")

> mpg\_ln <- fitdist(mpg, "lnorm")

> mpg\_w <- fitdist(mpg, "weibull")

> mpg\_plot.legend <- c("Gamma", "Lognormal", "Weibull")

> denscomp(list(mpg\_g, mpg\_ln, mpg\_w), legendtext = mpg\_plot.legend)



Conclusion –

We have categorised our variables appropriately according to the NOIR topology, and calculated a measure of central tendency from one of each type of variable.

Also for the two variables peak-rpm and city-mpg, we have estimated a distribution, namely normal/bimodal and gamma/lognormal respectively.