DESCRIPTIVE STATISTICS

Batch: 2021-6313

Presented by,

V. Jagadeesh

B. Tanuja Lakshmi

INDEX

1. INTRODUCTION

In the modern era, everything is now a data and data-driven system. Every second amount of data being generated is in terabytes. To draw meaningful statements and conclusions with numerical evidence, a widely used mathematics framework called **Statistics** is used.

So, what is meant by statistics?

- It means collection, organization, analysis and interpretation of data. Statistics are mainly used to give numerical conclusions.
- For example, if anyone asks you how many people are watching YouTube, in this case, we can't say more - many people are watching YouTube, we have to answer in numerical terms that give more meaning to you. We can say like during weekdays 6 pm-8 pm more people are watching YouTube applications and during weekend 8 pm-11 pm.
- If you want to answer active users, we can say there are two billion plus monthly active users, in the same way the users spend a daily average of 18 minutes. This is the numerical way to conclude the questions, and statistics is the medium used to make such inference.

Why do we have to learn Statistics?

We are learning statistics because we can observe the information properly and draw the conclusion from the large volume of the dataset, make reliable forecasts about business activity and improve the business process.

To do all kinds of these analyses, statistics are used. Further, it is classified into two types:

- Descriptive statistics
- Inferential statistics

Descriptive statistics summarize the data by computing mean, median, mode, standard deviation likewise. It is distinguished from inferential statistics by its aim to summarize the sample rather than use the data to learn more about the *Population* that sample of data thought to represent this means it is not developed based on probability theory.

Whereas inferential statistics are the methods for using sample data to make general conclusions (inferences) about populations by using the hypothesis. The sample is typically part of the whole population which contains only limited information about the population. For example, you might have seen the exit poll; those exit polls are calculated by taking several samples from different regions of that territory. Such conclusions are drawn from inferential statistics.

Statistics include:

- **Design of experiments:** Used to understand Characteristics of the dataset
- **Sampling:** Used to understand the samples
- **Descriptive statistics:** Summarization of data
- **Inferential Statistics:** Hypothesis way of concluding data
- **Probability Theory:** Likelihood estimation

2. DESCRIPTIVE STATISTICS

- Descriptive statistics summarize and organize characteristics of a data set. A data set is a collection of responses or observations from a sample or entire population.
- In quantitative research, after collecting data, the first step of statistical analysis is to describe characteristics of the responses, such as the average of one variable (e.g., age), or the relation between two variables (e.g., age and creativity).
- The next step is inferential statistics, which help you decide whether your data confirms or refutes your hypothesis and whether it is generalizable to a larger population.

In statistical analysis, there are three main fundamental concepts associated with describing the data:

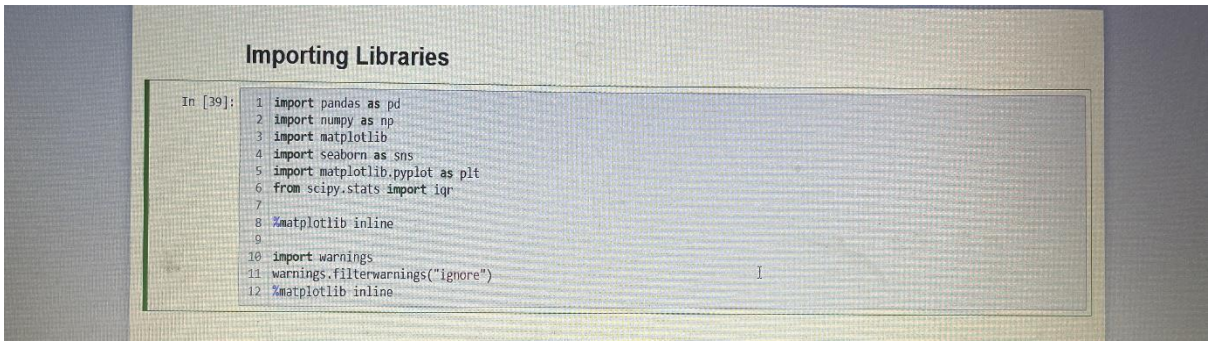
- location or Central tendency
- Dissemination or spread
- Shape or distribution.

A raw dataset is difficult to describe; descriptive statistics describe the dataset in a way simpler manner through;

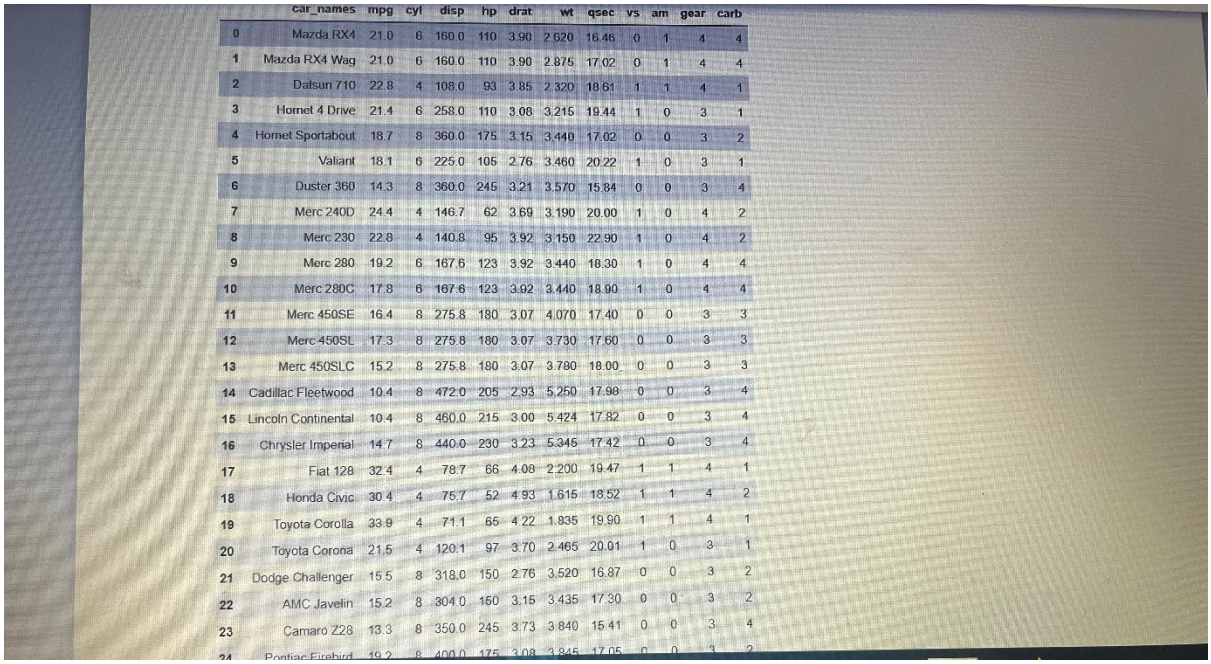
- The measure of central tendency (Mean, Median, Mode)
- Measure of spread (Range, Quartile, Percentiles, absolute deviation, variance and standard deviation)
- Measure of symmetry (Skewness)
- Measure of Peaked Ness (Kurtosis)

3. CODE IMPLEMENTATION: BASIC STATISTICS IN PYTHON

3.1 Importing Libraries

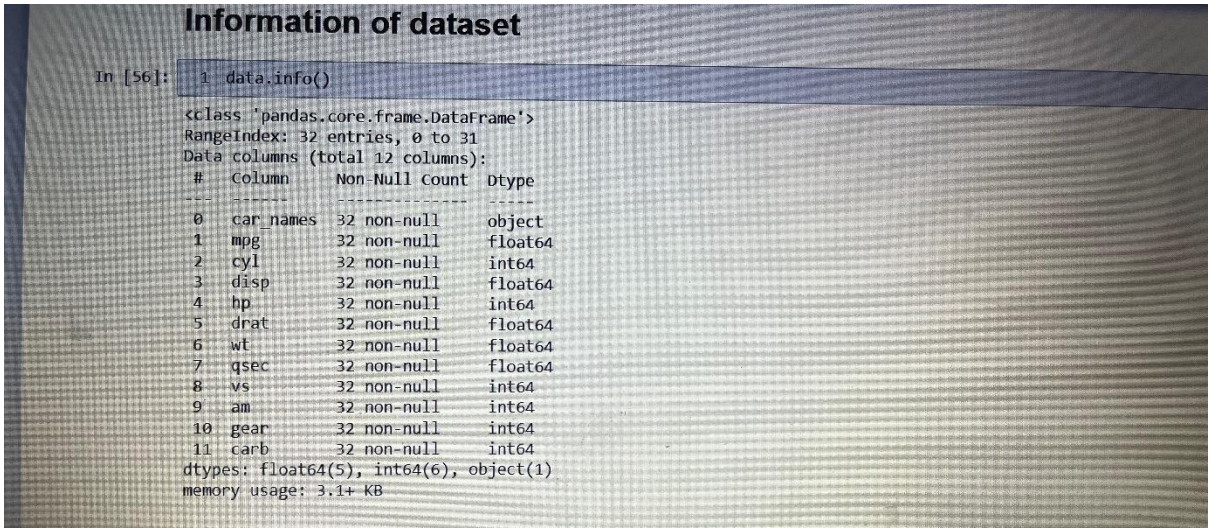


3.2 Read the dataset (mtcars)



	car_names	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
0	Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
1	Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
2	Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
3	Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
4	Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
5	Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
6	Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
7	Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
8	Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
9	Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
10	Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
11	Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
12	Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
13	Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
14	Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4
15	Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4
16	Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
17	Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
18	Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
19	Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1
20	Toyota Corona	21.5	4	120.1	97	3.70	2.465	20.01	1	0	3	1
21	Dodge Challenger	15.5	8	318.0	150	2.76	3.520	16.87	0	0	3	2
22	AMC Javelin	15.2	8	304.0	150	3.15	3.435	17.30	0	0	3	2
23	Camaro Z28	13.3	8	350.0	245	3.73	3.840	15.41	0	0	3	4
24	Pontiac Firebird	19.2	8	400.0	175	3.08	3.845	17.05	0	0	3	2

3.3 Information of dataset

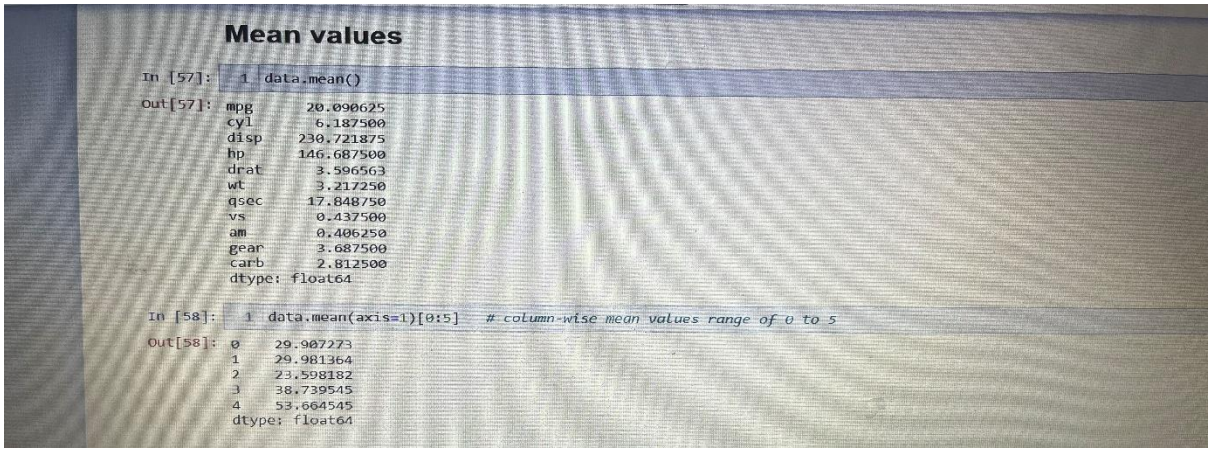


3.4 Measure of centre

- Measures of centre are statistics that give us a sense of the "middle" of a numeric variable.
- In other words, centrality measures give you a sense of a typical value you'd expect to see.
- Common measures of centre include the mean, median and mode.

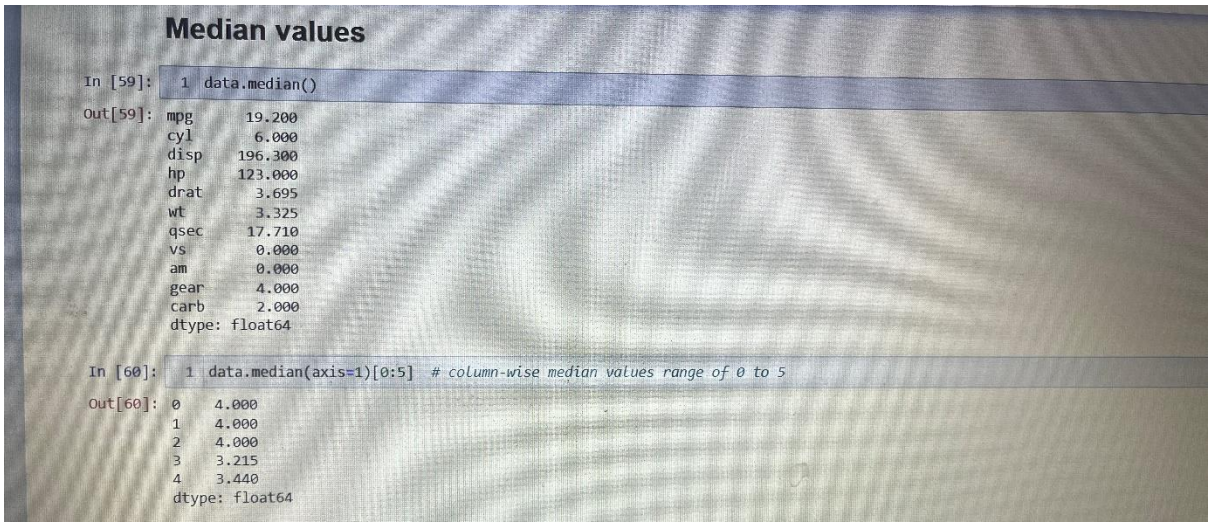
A. Mean

The average value of the data. Can be calculated by adding all the measurements of a variable together and dividing that summation by the number of observations used.



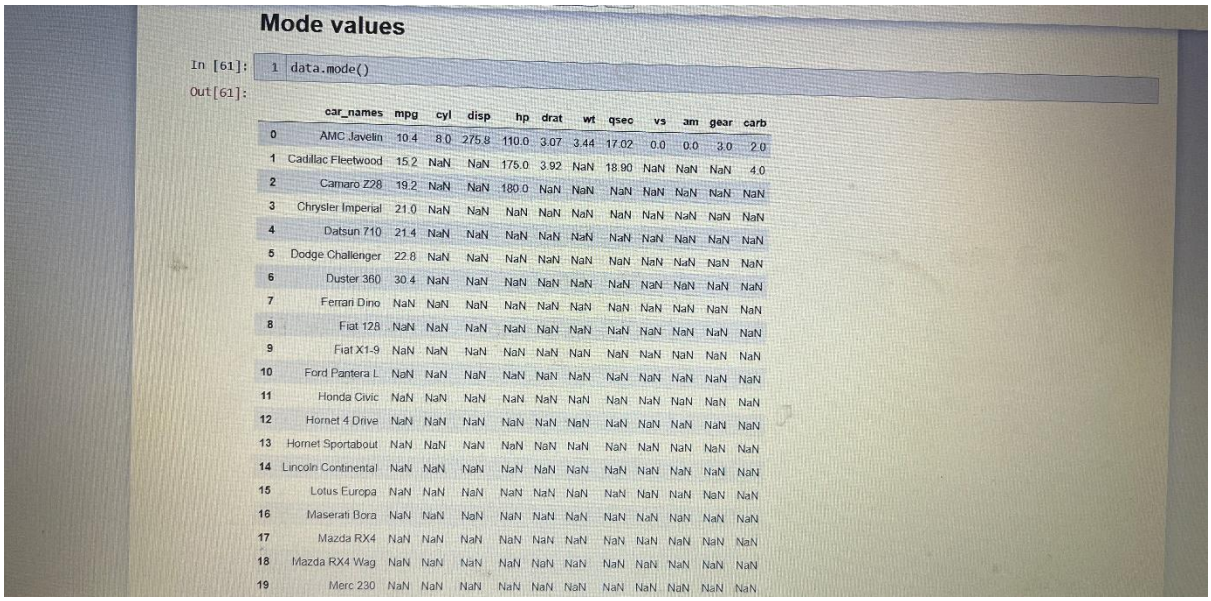
B. Median

The middle value when the measurements are placed in ascending order. If there is no true midpoint, the median is calculated by adding the two midpoints together and dividing by 2.



C. Mode

The number that occurs the most in the set of measurements.

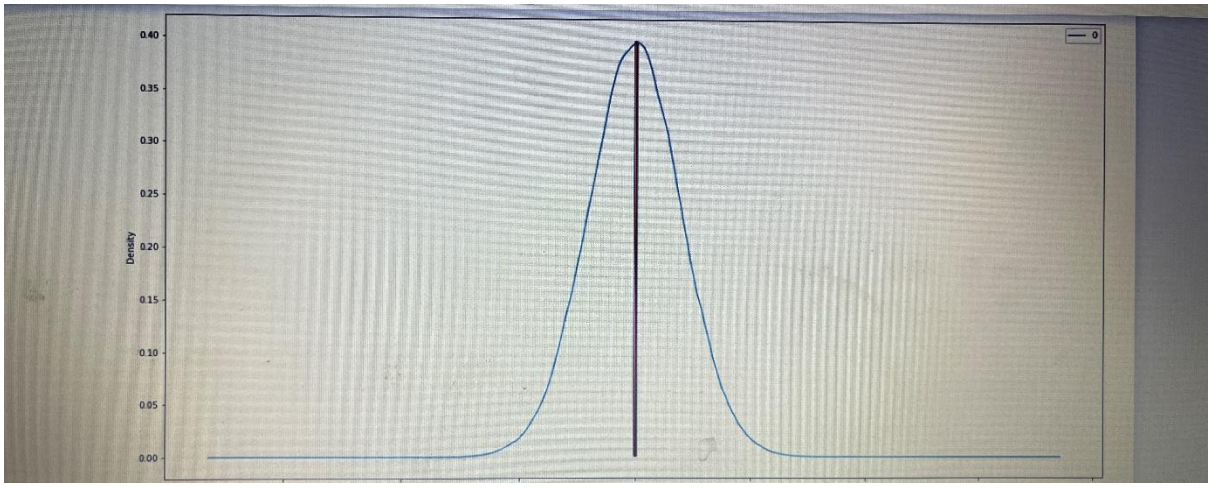


D. Density plot for mean and median

- Although the mean and median both give us some sense of the centre of a distribution, they aren't always the same.
- The median always gives us a value that splits the data into two halves.
- While the mean is a numeric average so extreme values can have a significant impact on the mean.

- In a symmetric distribution, the mean and median will be the same.
- In the plot the mean and median are both so close to zero that the red median line lies on top of the thicker black line drawn at the mean.

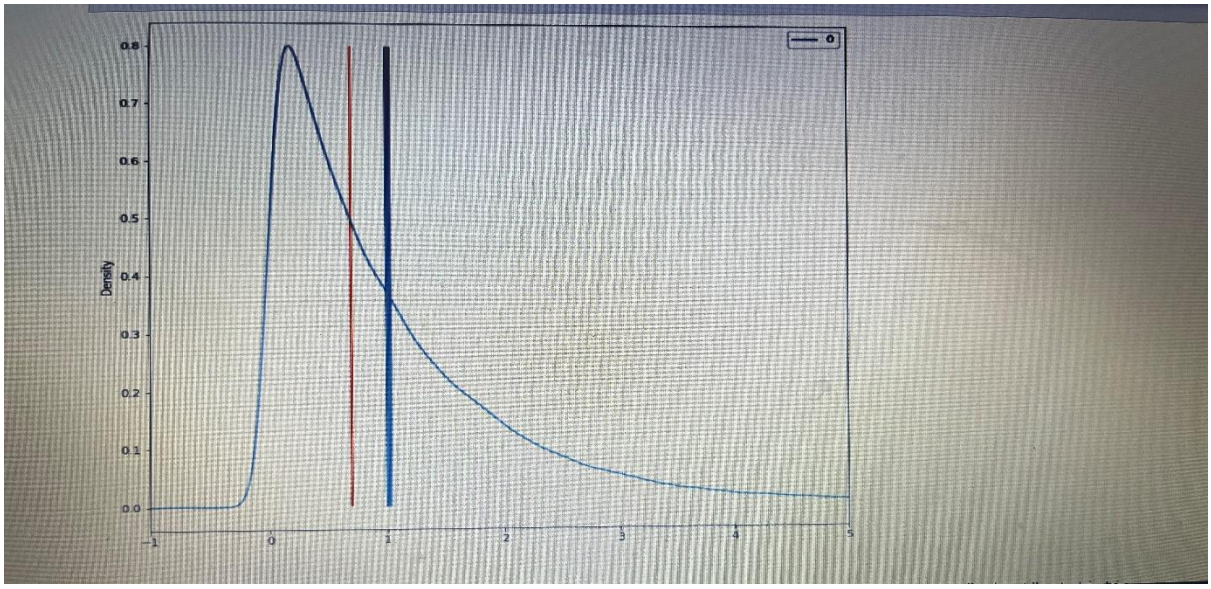
```
In [62]:
norm_data = pd.DataFrame(np.random.normal(size=100000))

norm_data.plot(kind="density",
               figsize=(20,10))


plt.vlines(norm_data.mean(),     # Plot black line at mean
           ymin=0,
           ymax=0.4,
           linewidth=5.0)

plt.vlines(norm_data.median(),   # Plot red line at median
           ymin=0,
           ymax=0.4,
           linewidth=2.0,
           color="red")
plt.show()
```



E. Skewed data distribution

- In skewed distributions, the mean tends to get pulled in the direction of the skew, while the median tends to resist the effects of skew.

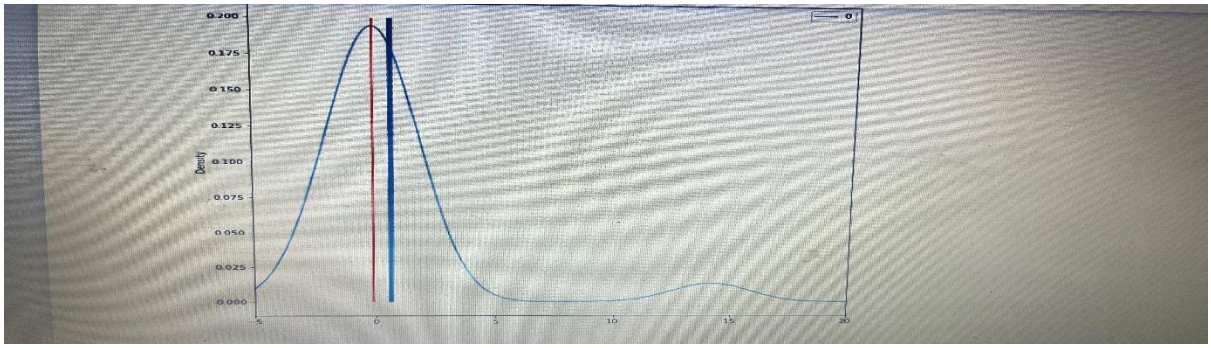
```
In [18]:
skewed_data = pd.DataFrame(np.random.exponential(size=100000))

skewed_data.plot(kind="density",
                 figsize=(10,10),
                 xlim=(-1,5))


plt.vlines(skewed_data.mean(),     # Plot blue line at mean
           ymin=0,
           ymax=0.8,
           linewidth=5.0)

plt.vlines(skewed_data.median(),   # Plot red line at median
           ymin=0,
           ymax=0.8,
           linewidth=2.0,
           color="red")

plt.show()
```



- Since the median tends to resist the effects of skewness and outliers, it is known a "robust" statistic.

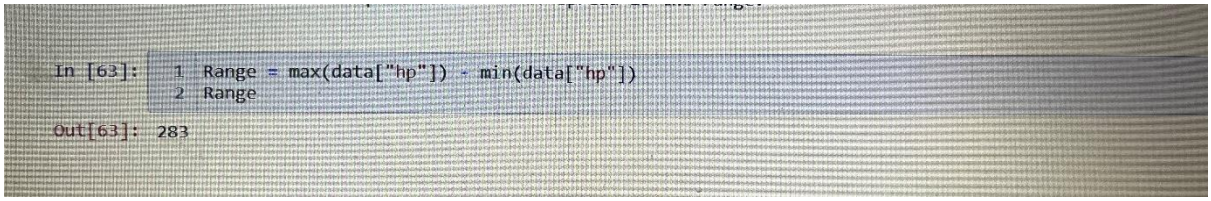
- The median generally gives a better sense of the typical value in a distribution with significant skew or outliers.

```
53]:  1 norm_data = np.random.normal(size=50)
      2 outliers = np.random.normal(15, size=3)
      3 combined_data = pd.DataFrame(np.concatenate((norm_data, outliers), axis=0))
      4
      5 combined_data.plot(kind="density",
      6                    figsize=(10,10),
      7                    xlim=(-5,20))
      8
      9
     10 plt.vlines(combined_data.mean(),     # Plot blue line at mean
     11            ymin=0,
     12            ymax=0.2,
     13            linewidth=5.0)
     14
     15 plt.vlines(combined_data.median(),   # Plot red line at median
     16            ymin=0,
     17            ymax=0.2,
     18            linewidth=2.0,
     19            color="red")
     20 plt.show()
```



3.5 Measure of Spread

- Measures of spread (dispersion) are statistics that describe how data varies.
- While measures of centre give us an idea of the typical value, measures of spread give us a sense of how much the data tends to diverge from the typical value.
- One of the simplest measures of spread is the range.

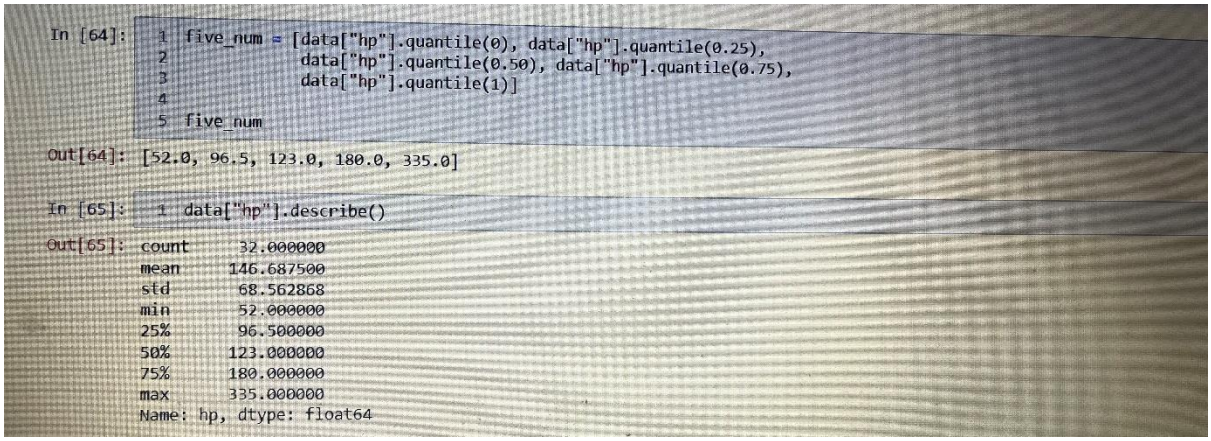


```
In [63]:  1 Range = max(data["hp"]) - min(data["hp"])
          2 Range

Out[63]: 283
```

A. Describing the specific column (hp)

- The median represents the 50 percentiles of a data set.
- A summary of several percentiles can be used to describe a variable's spread.
- We can extract data using the quantile () function

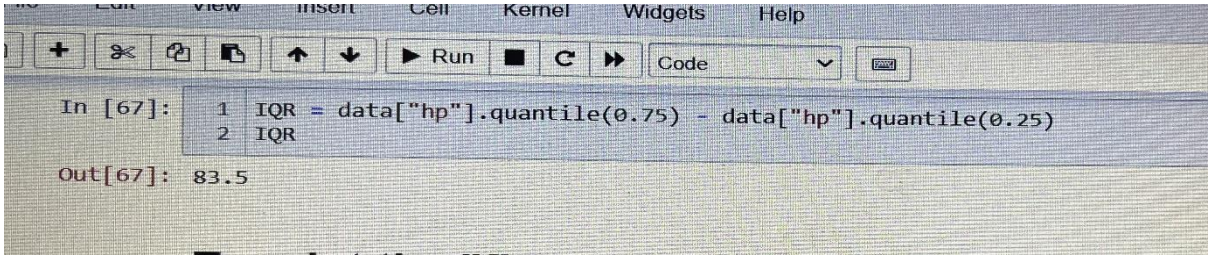
    1. Minimum value (0 percentile)
    2. First quartile (25 percentile)
    3. Median (50 percentile)
    4. Third quartile (75 percentile)
    5. Maximum value (100 percentile)



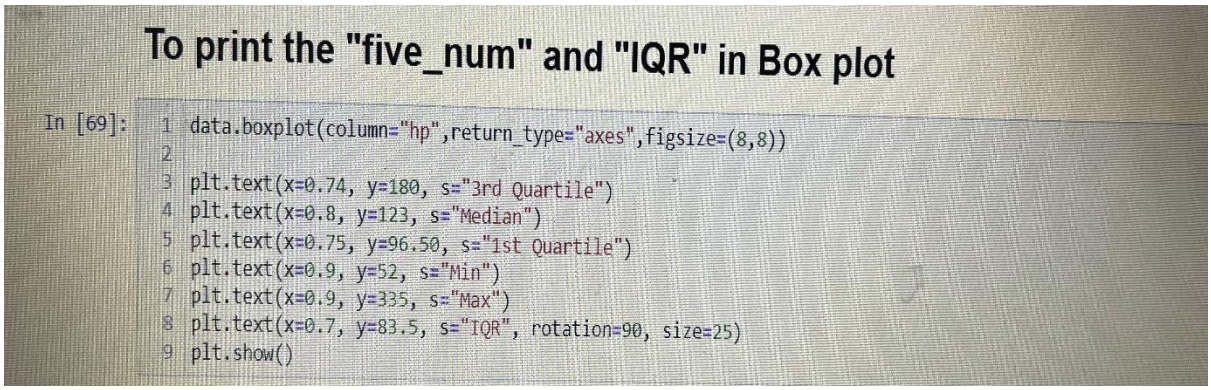
```
In [64]:  1 five_num = [data["hp"].quantile(0), data["hp"].quantile(0.25),
          2             data["hp"].quantile(0.50), data["hp"].quantile(0.75),
          3             data["hp"].quantile(1)]
          4
          5 five_num

Out[64]: [52.0, 96.5, 123.0, 180.0, 335.0]

In [65]:  1 data["hp"].describe()

Out[65]: count     32.000000
         mean     146.687500
         std       68.562868
         min       52.000000
         25%       96.500000
         50%      123.000000
         75%      180.000000
         max      335.000000
         Name: hp, dtype: float64
```

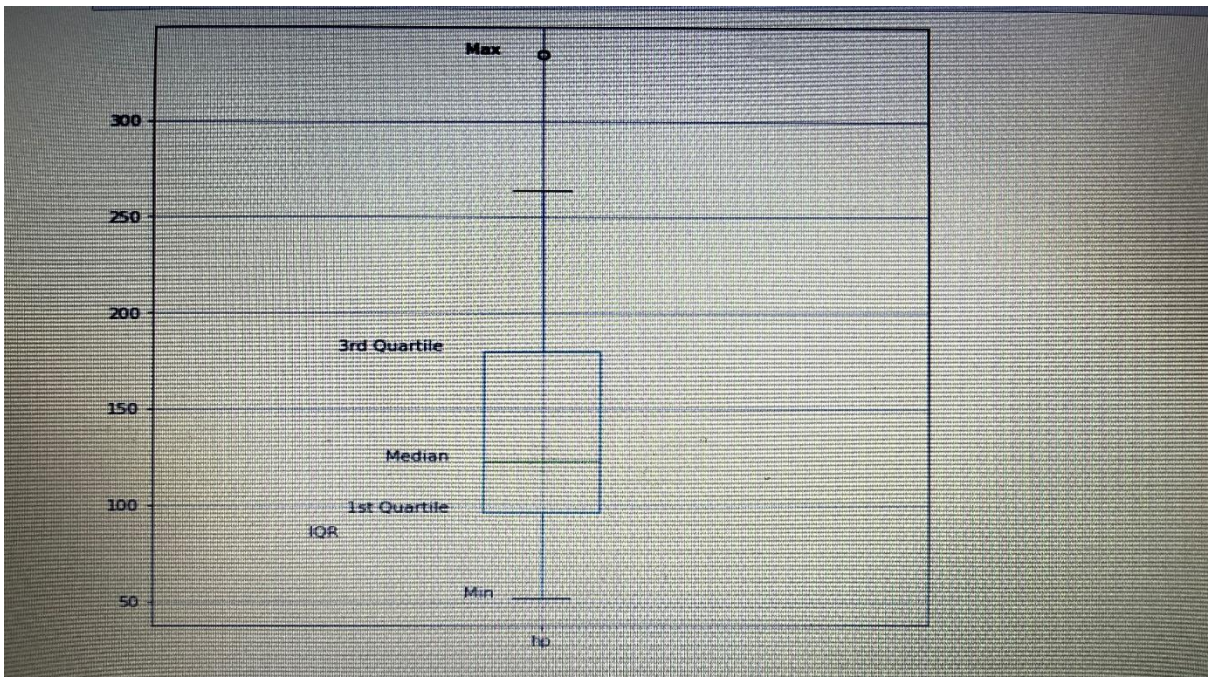
B. Inter Quantile Range (IQR)

- Interquartile (IQR) range is another common measure of spread.
- IQR is the distance between the 3rd quartile and the 1st quartile.
- Using specific column ("hp")



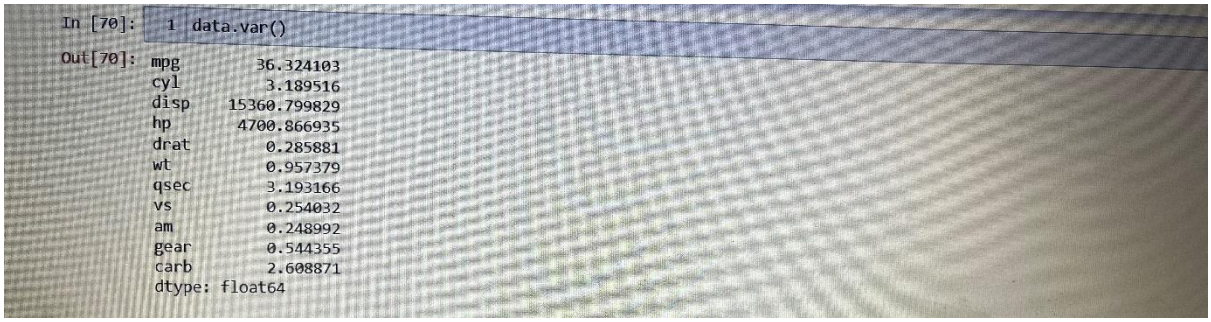
C. To print the "five_num" and "IQR" in Box plot





D. Variance

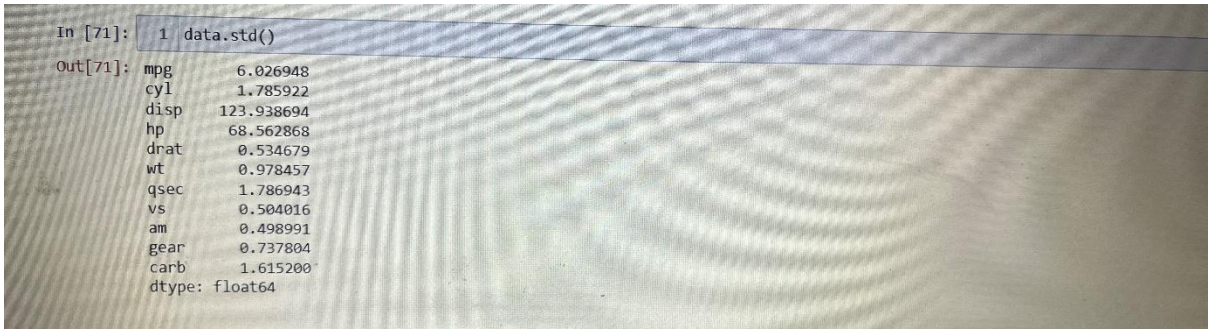
- Variance is another measure of dispersion.
- It is the square of the standard deviation and the covariance of the random variable with itself.
- The line of code below prints the variance of all the numerical variables in the dataset.
- The interpretation of the variance is similar to that of the standard deviation.



```
In [70]: 1 data.var()

Out[70]: mpg       36.324103
         cyl        3.189516
         disp   15360.799829
         hp      4700.866935
         drat       0.285881
         wt         0.957379
         qsec       3.193166
         vs         0.254032
         am         0.248992
         gear       0.544355
         carb       2.608871
         dtype: float64
```

E. Standard Deviation

- Standard deviation is a measure that is used to quantify the amount of variation of a set of data values from its mean.
- A low standard deviation for a variable indicates that the data points tend to be close to its mean, and vice versa.
- The line of code below prints the standard deviation of all the numerical variables in the data.



```
In [71]: 1 data.std()

Out[71]: mpg       6.026948
         cyl       1.785922
         disp    123.938694
         hp       68.562868
         drat      0.534679
         wt        0.978457
         qsec      1.786943
         vs        0.504016
         am        0.498991
         gear      0.737804
         carb      1.615200
         dtype: float64
```

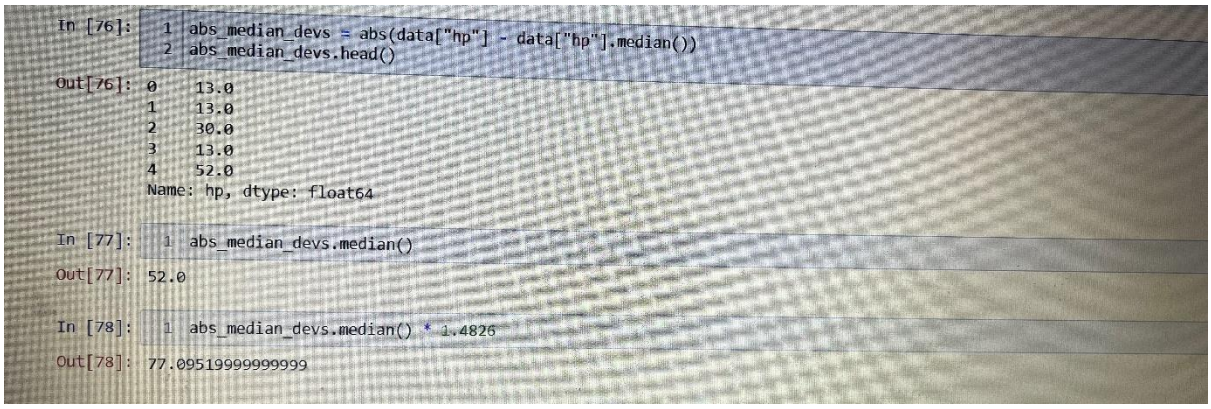
F. Median Absolute Deviation (MAD)

- Since variance and standard deviation are both derived from the mean and they are susceptible to the influence of data skew and outliers.

- Median absolute deviation is an alternative measure of spread based on the median, which inherits the median's robustness against the influence of skew and outliers.

- It is the median of the absolute value of the deviations from the median.

- The MAD is often multiplied by a scaling factor of 1.4826



G. Skewness and Kurtosis

- Beyond measures of centre and spread, descriptive statistics include measures that give you a sense of the shape of a distribution.

- Skewness measures the skew or asymmetry of a distribution while kurtosis measures how much data is in the tails of distribution vs the c entre.

- We won't go into the exact calculations behind skewness and kurtosis but they are essentially just statistics that take the idea of variance a step further.

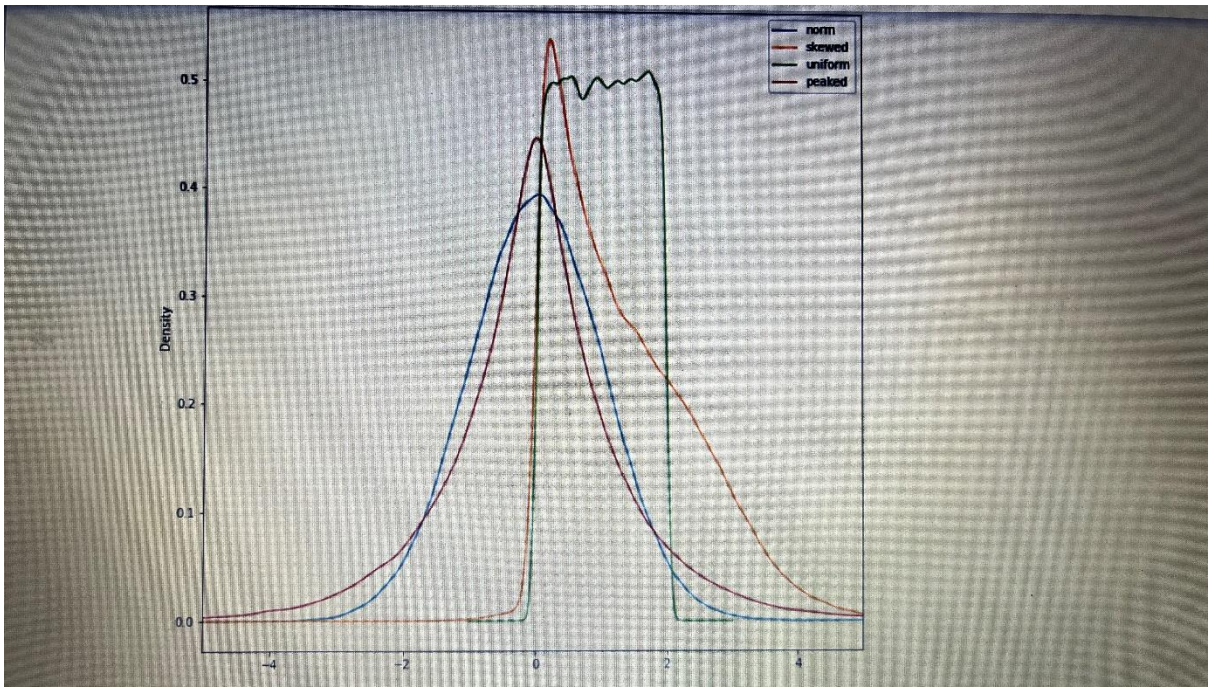
- While variance involves squaring deviations from the mean, skewness involves cubing deviations from the mean and kurtosis involves raising deviations from the mean to the 4th power.

```
In [79]:  1 data.skew()    # Check skewness
Out[79]: mpg     0.672377
         cyl    -0.192261
         disp    0.420233
         hp      0.799107
         drat    0.292780
         wt      0.465916
         qsec    0.406347
         vs      0.264542
         am      0.400809
         gear    0.582309
         carb    1.157091
         dtype: float64

In [80]:  1 data.kurt()    # Check kurtosis
Out[80]: mpg     -0.022006
         cyl      1.762794
         disp     1.067523
         hp       0.275212
         drat    -0.450432
         wt       0.416595
         qsec     0.864931
         vs      -2.063273
         am      -1.066596
         gear    -0.895292
         carb     2.020859
         dtype: float64
```

H. To explore these two measures - let's create some dummy data and inspect it by using Density Plot

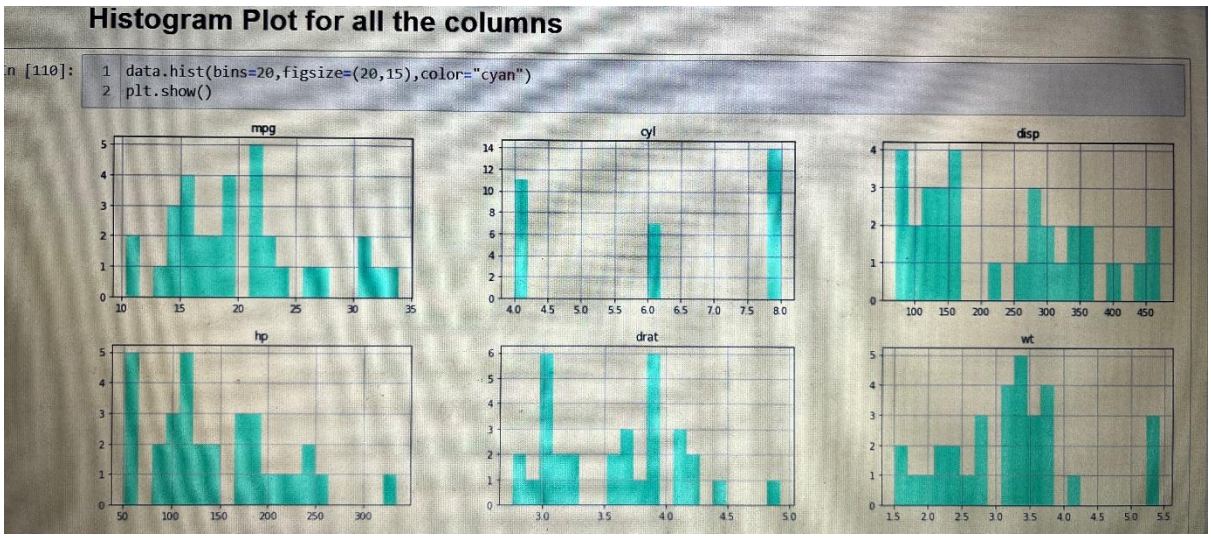
To explore these two measures - let's create some dummy data and inspect it by using Density Plot

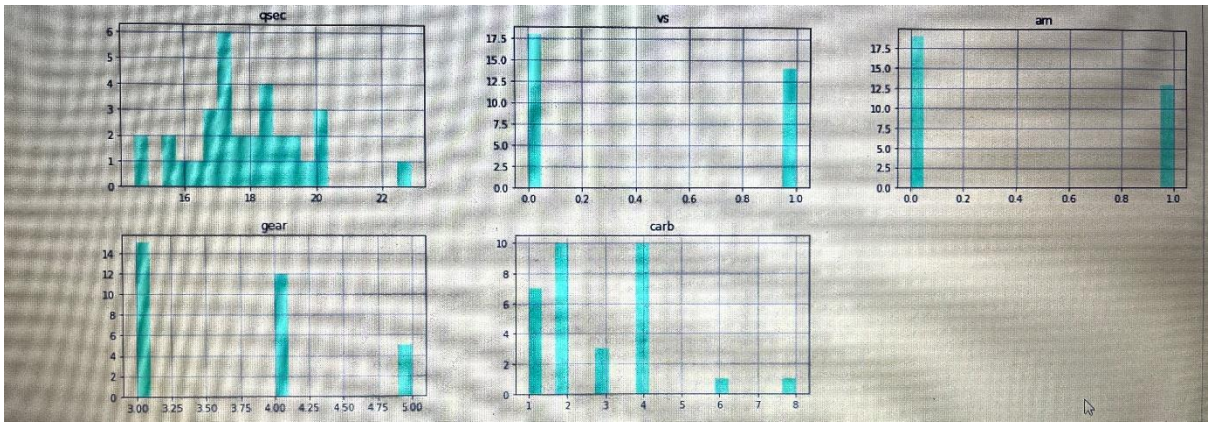
```
In [81]:
norm_data = np.random.normal(size=100000)
skewed_data = np.concatenate((np.random.normal(size=35000)+2,
                              np.random.exponential(size=65000)),
                              axis=0)
uniform_data = np.random.uniform(0,2, size=100000)
peaked_data = np.concatenate((np.random.exponential(size=50000),
                              np.random.exponential(size=50000)*(-1)),
                              axis=0)

data_df = pd.DataFrame({"norm":norm_data,
                        "skewed":skewed_data,
                        "uniform":uniform_data,
                        "peaked":peaked_data})

data_df.plot(kind="density",figsize=(10,10),xlim=(-5,5))
plt.show()
```



6. Histogram Plot for all the columns

- A histogram is used to summarize discrete or continuous data.
- In other words, it provides a visual interpretation. of numerical data by showing the number of data points that fall within a specified range of values called “bins”.
- Descriptive statistics enable you to compare various measures across the different variables.
- These include mean, mode, standard deviation, etc.



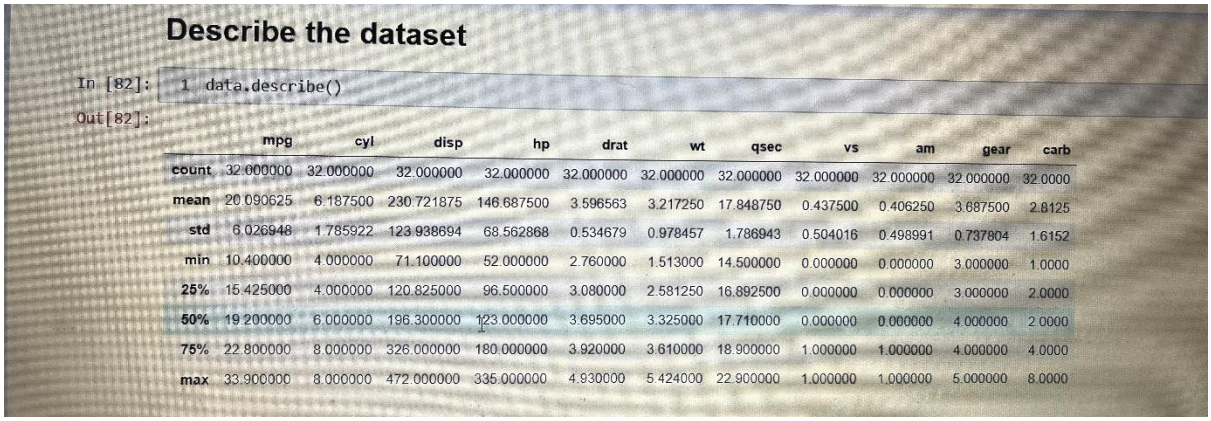


7. Describe the dataset

- The describe () method returns description of the data in the Data Frame.
- If the Data Frame contains numerical data, the description contains this information for each column:

    - count - The number of not-empty values
    - mean - The average (mean) value
    - std - The standard deviation
    - min - the minimum value
    - 25% - The 25% percentile
    - 50% - The 50% percentile
    - 75% - The 75% percentile
    - max - the maximum value

- Percentile meaning: how many of the values are less than the given percentile.



	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
count	32.000000	32.000000	32.000000	32.000000	32.000000	32.000000	32.000000	32.000000	32.000000	32.000000	32.0000
mean	20.090625	6.187500	230.721875	146.687500	3.596563	3.217250	17.848750	0.437500	0.406250	3.687500	2.8125
std	6.026948	1.785922	123.938694	68.562868	0.534679	0.978457	1.786943	0.504016	0.498991	0.737804	1.6152
min	10.400000	4.000000	71.100000	52.000000	2.760000	1.513000	14.500000	0.000000	0.000000	3.000000	1.0000
25%	15.425000	4.000000	120.825000	96.500000	3.080000	2.581250	16.892500	0.000000	0.000000	3.000000	2.0000
50%	19.200000	6.000000	196.300000	123.000000	3.695000	3.325000	17.710000	0.000000	0.000000	4.000000	2.0000
75%	22.800000	8.000000	326.000000	180.000000	3.920000	3.610000	18.900000	1.000000	1.000000	4.000000	4.0000
max	33.900000	8.000000	472.000000	335.000000	4.930000	5.424000	22.900000	1.000000	1.000000	5.000000	8.0000

CONCLUSION

Descriptive statistics help you explore features of your data, like centre, spread and shape by summarizing them with numerical measurements. It helps inform the direction of an analysis and let you communicate your insights to others quickly and succinctly. In addition, certain values, like the mean and variance, are used in all sorts of statistical tests and predictive models.

In this lesson, we generated a lot of random data to illustrate concepts, but we haven't actually learned much about the functions we've been using to generate random data. In the next lesson, we'll learn about probability distributions, including how to draw random data from them.

REFERENCES

1. https://analyticsindiamag.com/complete-guide-to-descriptive-statistics-in-python-for-beginners/

2. https://www.pluralsight.com/guides/interpreting-data-using-descriptive-statistics-python