Chapter 1. Exploratory data analysis

Classical statistics focused on drawing conclusions about large populations based on small samples. In 1962, John W. Tukey called for a reformation of statistics in his seminal paper “The Future of Data Analysis” and proposed a new scientific discipline called data analysis that included statistical inference as just one component and considered engineering and computer science communities.The field of exploratory data analysis was established with Tukey’s 1977 now-classic book Exploratory Data Analysis.

Elements of Structured Data

- EDA - In stat, it is an approach of analysing datasets to summarise their main characteristics, often using statistical graphics and other data visualisation methods.
- Data - Different types of information that usually is formatted in a particular manner.
- Data resources - sensor measurements, text, images, videos, data from IOT devices (real time data) like automation data, location data, status data etc.

Data must be put into a structured format for us to statistically analyse it.Unstructured data can be broken down into couple of forms:

- Numerical
  - Continuous (interval, float, numeric)
  - Discrete(integer, count)
- Categorical
  - Binary(logical, indicator, boolean)
  - Ordinal (order based-small-smaller-smallest)

Data Structures

Rectangular Data

Matrix of records with column denoting variables and rows denoting records.Data doesn’t always comes in various forms it’s important that we extract and transform data into rectangular.

Terminologies:

1.DataFrame:The format in which data will be loaded, generally in python or R.

2.Indexes:Index of the data frame to increase efficiency while working on data operations.

3..Features:Columns in data table.

4.Target: Output variable of data.

5.Records:Number of rows in the data.

Tools: Excel, spreadsheets, database table.

<u>Non - Rectangular Data</u>

Time series data measures successive measurements of the same variable.

Spatial data- focuses on spatial coordinates like pixels.

Text data -data is counted as words or alphabets.

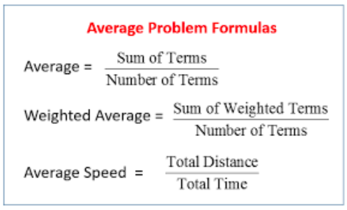
Image data.

Estimates of Location

Getting a “typical value” for each feature.

An estimate of where most of the data is located also called central tendency.

Mean :Sum of all values divide by number of values.



Average Problem Formulas

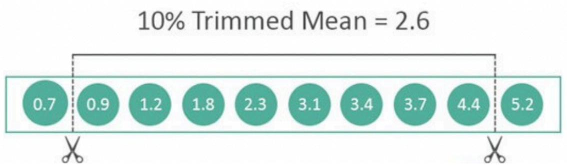
$$\text{Average} = \frac{\text{Sum of Terms}}{\text{Number of Terms}}$$

$$\text{Weighted Average} = \frac{\text{Sum of Weighted Terms}}{\text{Number of Terms}}$$

$$\text{Average Speed} = \frac{\text{Total Distance}}{\text{Total Time}}$$

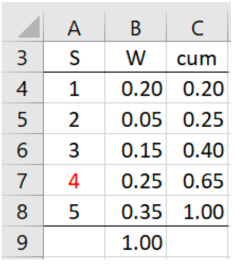
Weighted Mean :Sum of all values times a weight divided by the sum of the weights.Same as mean, expect you multiply every value by some xi before adding them up and dividing by the number of instances.

Trimmed Mean :The average of all values after dropping a fixed number of extreme values.Helps reduce the impact of outliers.



Median :The value in the sorted data such that one-half of it lies above and other below it.

Weighted Median :The value such that one-half of the sum of the weights lies above and other below the sorted data.



	A	B	C
3	S	W	cum
4	1	0.20	0.20
5	2	0.05	0.25
6	3	0.15	0.40
7	4	0.25	0.65
8	5	0.35	1.00
9		1.00	

1. column A - data ,column B - weights, column 3 - cumulative sum; sort the data in ascending order and then create a cumulative sum of the weights.

2. in column C whose value is larger than .50, the corresponding value in cell A is the weighted median.

Robust : Metric which is not sensitive to extreme values. Eg: median, weighted median, trimmed mean.

Outliers :A data value that is different from most of the data points.

Estimates of Variability(Dispersion Metrics)

Estimate of location tells us the centres of the data while variability tells spread of data by summing up it in a single number considering the central tendency.Variability is at the heart of statistics, where a lot of information on a dataset can be cleaned.

Derivations :The difference between observed values and the estimate of location also called errors or residuals.

Variance : Square the derivations, from the mean, divide by (n-1) where n is the number of instances.

Standard deviation : Square root of the variance.

Mean absolute deviation : Mean of the absolute value of the derivation from the mean, also known as l1-norm or Manhattan Norm.It is simply the sum of the absolute values of all deviations(values- mean) divided by the number of instances.

Median absolute deviation from the median : Median of the absolute values of the deviation from the median.

Range : Difference between the largest and smallest values in the dataset.

Order statistics :Metrics based on the data values sorted from smallest to largest.

Percentile :The value such that P percent of the data lies below it.(10th, 20th…,90th percentiles know as deciles)

Interquartile Range :Difference between the 75th and 25th percentile.

Note:

1. Reason to prefer std dev to the variance is that it operates on the same scale as the rest of the data.

2. Mean absolute deviation, std dev, variance and range are not robust against outliers.

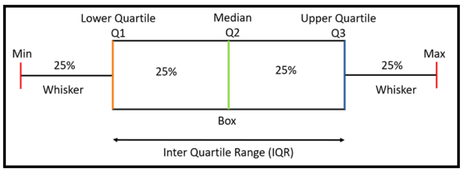
3. Median absolute deviation, interquartile range is robust against outliers.

4. For very large data sets, it can get computationally expensive to calculate these stats.since it requires sorting all the data values, so machine learning and statistical software algorithms are used.

Exploring the Data Distribution

Percentiles and Box-plot

Percentiles are a great way to summarise the tails of a distribution such as the top 1%.Boxplots are based on percentiles to see the central data, with top and bottom of the box are the 75th and 25th percentiles.Median is shown by the horizontal line.dashed lines are called whiskers which shows the range of data.An outlier is any number less than Q1−(1.5×IQR) or greater than Q3+(1.5×IQR).



Frequency Table and Histograms

Frequency table of a variable divides up the variable range into equally spaced segments and quantify the numbers of observations in each segment.

A histogram is a way to visualise a frequency table with bins on the x-axis and data count on the y-axis. Histograms are plotted such that:

1. Empty bins are included in the graph.

2. Bins are equal width.

3. Number of bins (or, equivalently, bin size) is up to the user.

4. Bars are contiguous — no empty space shows between bars, unless there is an empty bin.

- **Statistical Moments**

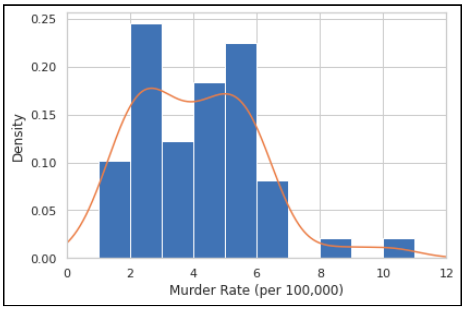
  First Moment is Location (Mean): Average of all the data.

  Second Moment is variability: How closely values are spread around the mean (or another measure of central tendency).

Third Moment is Skewness: Direction of the tail of the data - discovered through visualisation not a metric.

Fourth Moment is kurtosis: propensity for data having extreme values - discovered through visualisation not a metric.

Density plot



Smoothed histogram which shows the distribution of data values as a continuous line. It corresponds to plotting the histogram as a proportion rather than counts.

Exploring Binary and Categorical Data

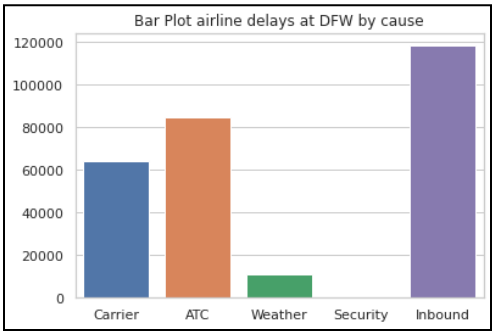
Mode: Simple summary statistic for categorical data which represents most commonly occurring category or value in a data set.

Expected value: When the categories can be associated with a numeric value, this gives an average value based on a category's probability of occurrence.(form of weighted mean)

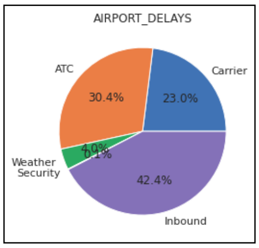
The expected value is calculated as follows:
1. Multiply each out come by its probability of occurring.
2. Sum these values.

$$EV = (0.05)(300) + (0.15)(50) + (0.80)(0) = 22.5$$

Bar plot: The frequency or proportion for each categorical variable plotted as bars with categories on the x-axis and frequencies or proportions on the y-axis.In a bar chart, the bars are shown separate from one another.



Pie chart: The frequency or proportion for each category plotted as wedges in a pie.



Exploring Two or More Variables

Mean and variance are a form of univariate analysis, correlation is a form of bivariate analysis and methods that look at more than two variables simultaneously are called multivariate analysis. After analysing individual variables you'll usually want to see how different variables interact with one another.When deciding which method to use, you must first determine whether you're comparing numerical variables to one another, numerical vs categorical variables, or categorical variables to one another.

- Numeric Vs Numeric Data

Correlation

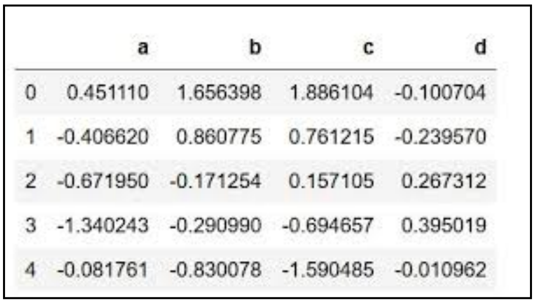
Correlation plot: is commonly plotted to visually display the relationship between multiple variables.Methods used to determine correlations:

1. Scatter plot

2. K-Pearson coefficient correlation

3. Spearman Rank.

Correlation matrix



	a	b	c	d
0	0.451110	1.656398	1.886104	-0.100704
1	-0.406620	0.860775	0.761215	-0.239570
2	-0.671950	-0.171254	0.157105	0.267312
3	-1.340243	-0.290990	-0.694657	0.395019
4	-0.081761	-0.830078	-1.590485	-0.010962

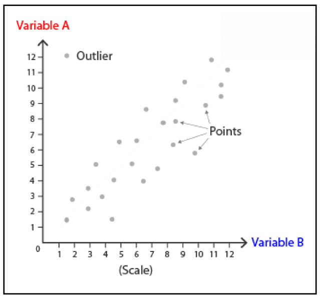
Correlation coefficient: A metric that measures the extent to which numeric variables are associated with one another (ranges from –1 (perfect negative correlation) to +1(perfect positive correlation)and 0 indicates no correlation).Generally use Pearson’s correlation coefficient.The PCC is sensitive to outliers.

Correlation matrix: A table which shows the correlation between all the variables and it shows on both rows and columns, and the cell values are the correlations between the variables. The primary diagonal will be 1’s the entire way down.

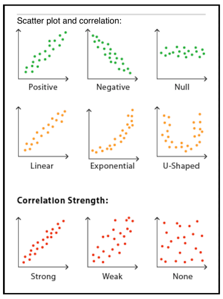
Scatterplot

It plots one numerical variable on one axis vs another on a different axis with each point correlating to a record.Scatterplots offers an excellent way to measure the relationship between two variables. You can use scatterplot against all the variables in your dataset to reduce relationship that would not otherwise be apparent.

Anatomy of scatterplot



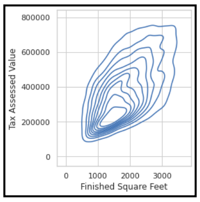
Scatter plot relation types



Hexagonal Binning: If you have a large number of values then scatter plots becomes plots become too hectic to gain insights from, One major benefit of hexagonal binning is that, by binning the data, the number of points that your computer has to render goes down substantially and offers speed benefits and reduces the size of jupyter notebooks.

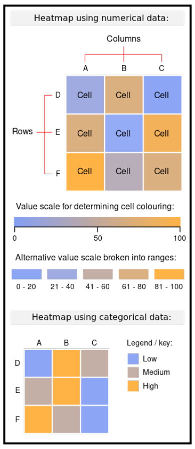
Contour plot: It is overlaid onto scatterplot and offers a layer

of lines that help determine where data is the most dense.



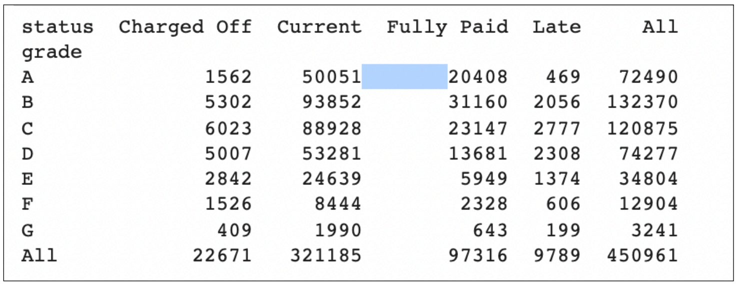
Heat Maps:

It offers another way to look at data using colours to communicate densities.They can also be used with categorical variables.


Heatmap using numerical data:
Columns
A
B
C
Rows
D
E
F
Cell
Cell
Cell
Cell
Cell
Cell
Cell
Cell
Cell
Value scale for determining cell colouring:
0
50
100
Alternative value scale broken into ranges:
0 - 20
21 - 40
41 - 60
61 - 80
81 - 100
Heatmap using categorical data:
A
B
C
D
E
F
Legend / key:
Low
Medium
High

- Categorical Vs Categorical Data

Contingency tables count the number of instances of two variables within the dataset.You can also use other metrics like the proportion and percent of each variable combination you want to as well.

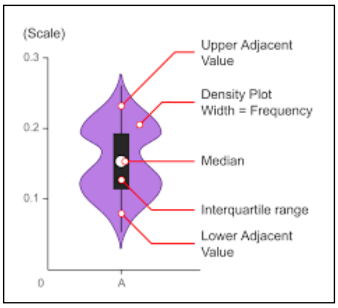


```
status  Charged Off  Current  Fully Paid  Late     All
grade
A              1562    50051       20408   469   72490
B              5302    93852       31160  2056  132370
C              6023    88928       23147  2777  120875
D              5007    53281       13681  2308   74277
E              2842    24639        5949  1374   34804
F              1526     8444        2328   606   12904
G               409     1990         643   199    3241
All           22671   321185       97316  9789  450961
```

- Categorical Vs Numerical Data

Categorical and numerical variables can be compared through, box plot,

Violin plot: It is variation of box plot which shows the distribution of values within the boxes.This to see where the density of values biases.



Visualizing multiple Variables

Compare multiple variables to one another using Plotly library, you can use a Facet or Trellis plot which essentially lines up different variables on the same X and Y axes plots them on multiple plots.