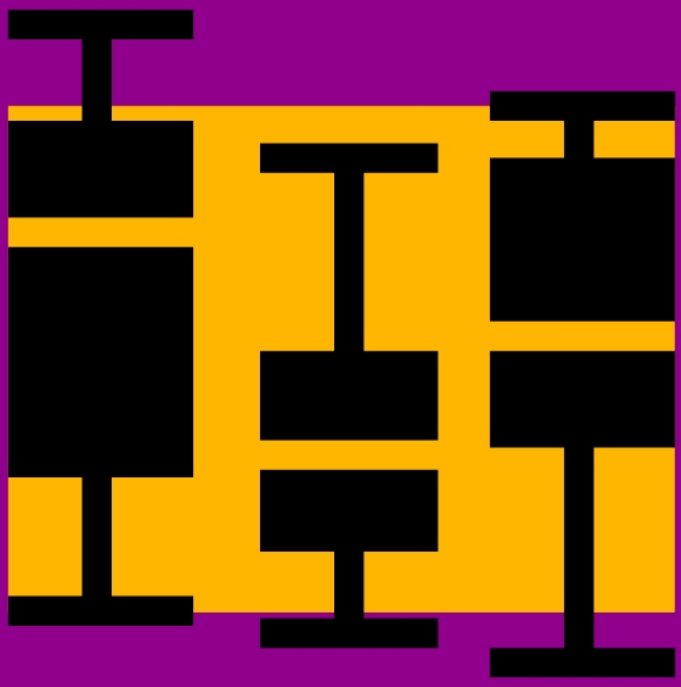
Exploratory Data Analysis



Rad Agustin

:)

Contents

1 Introduction **1**
1.1 Overview . . . 1
1.1.1 Approach . . . 1
1.1.2 Focus . . . 1
1.1.3 Philosophy . . . 1
1.1.4 Techniques . . . 2
1.1.5 How Does Exploratory Data Analysis differ from Classical Data Analysis? 2
1.1.6 The Role of Graphics . . . 3
1.2 Key Concepts . . . 5
1.2.1 Data Exploration . . . 5
1.2.2 Data Visualization . . . 5
1.2.3 Summary Statistics . . . 5

2 Data Collection and Cleaning **7**
2.1 Sources of Data and Data Types . . . 7
2.1.1 Classifications and Sources of Data . . . 7
2.1.2 General Classification of Collecting Data . . . 7
2.1.3 Data Types . . . 8
2.2 Data Cleaning and Preprocessing Techniques . . . 9
2.2.1 Common Data Cleaning Techniques . . . 9
2.2.2 Data Processing . . . 10
2.2.3 Dealing with Missing Values and Outliers . . . 11
2.2.4 Methods to Treat Missing Values . . . 12
2.2.5 Techniques of Outlier Detection and Treatment . . . 14
2.2.6 Dealing with Outliers . . . 15

2.2.7 Retain Outliers 16
2.2.8 Remove Outliers 16

3 Descriptive Statistics **17**
3.1 Frequencies and Descriptive Statistics 17
3.1.1 Histograms 19
3.1.2 Boxplots 20
3.1.3 Significance of Descriptive Statistics 22

4 Data Visualization **23**
4.1 Presentation of Data 23
4.1.1 Textual Presentation 23
4.1.2 Tabular Presentation 23
4.1.3 Graphical Presentation 25

5 EDA Techniques **29**
5.1 Correlation and Covariance 29
5.1.1 What is Covariance? 29
5.1.2 What is Correlation? 30
5.1.3 Difference Between Correlation and Covariance 31
5.2 Re-expressing Data 31
5.2.1 The log Transformation 32
5.2.2 The Tukey Transformation 32
5.2.3 Box-Cox Transformation 32
5.3 Median Polish 33
5.3.1 Data Structure 33
5.3.2 Decomposition 33
5.3.3 Robustness 33
5.3.4 Interpretation 34
5.3.5 Applications 34

6 References **35**

Chapter 1
Introduction

Section 1.1: Overview

1.1.1 Approach

Exploratory data analysis or "EDA" is a critical first step in analyzing the data from an experiment. Here are the main reasons we use EDA:

1. detection of mistakes
2. checking of assumptions
3. preliminary selection of appropriate models
4. determining relationships among the explanatory variables, and
5. assessing the direction and rough size of relationships between explanatory and outcome variables.

1.1.2 Focus

The EDA approach is precisely not a set of techniques, but an attitude/philosophy about how a data analysis should be carried out.

1.1.3 Philosophy

EDA is not identical to statistical graphics although the two terms are used almost interchangeably. Statistical graphics is a collection of techniques - all graphically based and all focusing on one data characterization aspect, whereas:

- EDA encompasses a larger venue; EDA is an approach to data analysis that postpones the usual assumptions about what kind of model the data follow with the more direct approach of allowing the data itself to reveal its underlying structure and model.

- EDA is not a mere collection of techniques; EDA is a philosophy as to how we dissect a data set; what we look for; how we look; and how we interpret. It is true that EDA heavily uses the collection of techniques that we call "statistical graphics", but it is not identical to statistical graphics per se.

1.1.4 Techniques

Most EDA techniques are graphical in nature with a few quantitative techniques. The reason for the heavy reliance on graphics is that by its very nature the main role of EDA is to open-mindedly explore, and graphics gives the analysts unparalleled power to do so, enticing the data to reveal its structural secrets, and being always ready to gain some new, often unsuspected, insight into the data. In combination with the natural pattern-recognition capabilities that we all possess, graphics provides, of course, unparalleled power to carry this out.

The particular graphical techniques employed in EDA are often quite simple, consisting of various techniques of:

1. Plotting the raw data (e.g., data traces, histograms, probability plots)
2. Plotting simple statistics (e.g, mean plots, standard deviation plots, box plots)
3. Positioning such plots so as to maximize our natural pattern-recognition abilities, such as using multiple plots per page.

1.1.5 How Does Exploratory Data Analysis differ from Classical Data Analysis?

Three popular data analysis approaches are:

1. Classical
2. Exploratory (EDA)
3. Bayesian

These three approaches are similar in that they all start with a general science/engineering problem and all yield science/engineering conclusions. The difference is the sequence and focus of the intermediate steps.

For classical analysis, the sequence is
$Problem \rightarrow Data \rightarrow Model \rightarrow Analysis \rightarrow Conclusions$

For EDA, the sequence is
$Problem \rightarrow Data \rightarrow Analysis \rightarrow Model \rightarrow Conclusions$

For Bayesian, the sequence is
$Problem \rightarrow Data \rightarrow Model \rightarrow PriorDistribution \rightarrow Analysis \rightarrow Conclusions$

Thus, for classical analysis, the data collection is followed by the imposition of a model (normality, linearity, etc.) and the analysis, estimation, and testing that follows are focused on the parameters of that model.

For EDA, the data collection is not followed by a model imposition; rather it is followed immediately by analysis with a goal of inferring what model would be appropriate.

Finally, for a Bayesian analysis, the analyst attempts to incorporate scientific/engineering knowledge/expertise into the analysis by imposing a data- independent distribution on the parameters of the selected model; the analysis thus consists of formally combining both the prior distribution on the parameters and the collected data to jointly make inferences and/or test assumptions about the model parameters.

The primary goal of EDA is to maximize the analyst's insight into a data set and into the underlying structure of a data set, while providing all of the specific items that an analyst would want to extract from a data set, such as:

- a good-fitting, parsimonious model
- a list of outliers
- a sense of robustness of conclusions
- estimates for parameters
- uncertainties for those estimates
- a ranked list of important factors
- conclusions as to whether individual factors are
- statistically significant
- optimal settings

1.1.6 The Role of Graphics

Statistics and data analysis procedures can broadly be split into two parts:

1. Quantitative

2. Graphical

Quantitative techniques are the set of statistical procedures that yield numeric or tabular output. Examples of quantitative techniques include:

- hypothesis testing
- analysis of variance
- point estimates and confidence intervals
- least squares regression

These and similar techniques are all valuable and are mainstream in terms of classical analysis.

On the other hand, there is a large collection of statistical tools that we generally refer to as ***graphical techniques***. These include:

- scatter plots
- histograms
- probability plots
- residual plots
- boxplots
- block plots

EDA Approach Relies Heavily on Graphical Techniques

Graphical procedures are not just tools that we could use in an EDA context, they are tools that we must use. Such graphical tools are the shortest path to gaining insight into a data set in terms of

- Testing assumptions
- model selection
- model validation
- estimator selection
- relationship identification
- factor effect determination
- outlier detection

If one is not using statistical graphics, then one is forfeiting insight into one or more aspects of the underlying structure of the data.

Section 1.2: Key Concepts

1.2.1 Data Exploration

Data exploration is a statistical process that lets you see how your data is distributed, identify any outliers, and determine which statistical tests might be most appropriate. When deciding what type of analysis or interpretation is most accurate for your data, preliminary data exploration can help you understand its characteristics.

1.2.2 Data Visualization

Data visualization is the representation of data through use of common graphics, such as charts, plots, infographics and even animations. These visual displays of information communicate complex data relationships and data-driven insights in a way that is easy to understand.

1.2.3 Summary Statistics

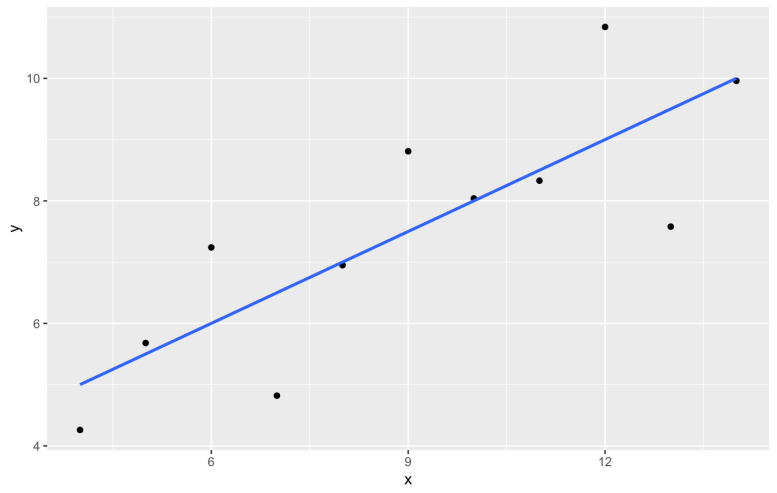
Data summaries use descriptive (summary) statistics to present collected research data in a logical, meaningful, and efficient way. In most cases, data summaries do not make inferences about the data and its ability to prove or disprove a research question.

Data summaries usually present the dataset's average (mean, median, and/or mode); standard deviation from mean or interquartile range; how the data is distributed across the range of data (for example is it skewed to one side of the range); and statistical dependence (if more than one variable was captured in the dataset). Data summaries may be presented in numerical text and/or in tables, graphs, or diagrams.

Example:

X	Y
10.00	8.04
8.00	6.95
13.00	7.58
9.00	8.81
11.00	8.33
14.00	9.96
6.00	7.24
4.00	4.26
12.00	10.84
7.00	4.82
5.00	5.68

N	11
Mean of X	9.0
Mean of Y	7.5
Intercept	3
Slope	0.5
Residual Std Dev	1.237
Correlation	0.816



- The data set “behaves like” a linear curve with some scatter;
- there is no justification for a more complicated model (e.g., quadratic);
- there are no outliers;
- the vertical spread of the data appears to be of equal;
- height irrespective of the X-value; this indicates that the data are equally-precise throughout and so a “regular” (that is, equi-weighted) fit is appropriate.

Chapter 2
Data Collection and Cleaning

Section 2.1: Sources of Data and Data Types

2.1.1 Classifications and Sources of Data

1. Primary vs Secondary
    - Primary source - data measured by the researcher/agency that publish it
    - Secondary source – any republication of data by another agency

    Example: The publications of the Philippine Statistics Authority (PSA) are primary sources and all subsequent publications of other agencies are secondary sources.
2. External vs Internal
    - Internal data – information that relates to the operations and functions of the organization collecting the data
    - External data – information that relates to some activity outside the organization collecting the data

    Example: The sales data of SM is internal data for SM but external for any other organization such as Robinson's and Ayala

2.1.2 General Classification of Collecting Data

1. By census
2. By survey sampling

Census or ***complete enumeration*** is the process of gathering information from every unit in the population.

Advantages of taking census:

- Sometimes census makes more sense than using samples.
- It eliminates the chance of randomly selecting samples that might not be representative of the population.
- Clients feel more comfortable to census than sampling.

Disadvantages of taking census:

- It is not always possible to get timely, accurate and economical data.
- It is costly especially if the number of units in the population is too large.

Survey sampling is the process of obtaining information from the units in the selected sample.

Advantages of Survey Sampling:

- Reduced cost – sampling can save money
- Greater speed – sampling can save time
- Greater scope – for given resources, the sample can broaden the scope of the study
- Greater accuracy – if assessing the population is impossible, the sample is the only option

2.1.3 Data Types

1. Qualitative Data
    - a data that yields categorical responses

   Example: Political affiliation, occupation, marital status, religion, sex
2. Quantitative Data
    - takes on numerical values representing an amount or quantity

   Example: Weight, height, number of cars, age, number of dependents

Types of Quantitative Data

1. Discrete Data
    - assumes finite, or, at most, countably infinite number of values; usually measured by counting or enumeration.

Examples: number of dependents, number of cars
If we let X be the number of dependents, then $X = 3$ if there are 3 dependents and $X = 1$ if there is only one dependent.
If we let Y be the number of cars owned by residents in a subdivision, then $Y = 2$ if there are 2 cars owned by resident A and $Y = 4$ if there are 4 cars owned by resident B.

2. Continuous Data

   - assumes infinitely many values corresponding to a line interval. It is measurable (measured using a continuous scale such as kilos, centimeters, grams)

   Examples: weight, height
   Weight of a person may not exactly 60 kilos. Maybe it is 60.1 kilos or 60.2 kilos or 59.9 kilos or 59.8 kilos. Thus, one's weight may be in the interval $(59.5, 60.5)$ kilos. If we let X be the weight of a person, the interval $(59.5, 60.5)$ can be written as $59.5 < X < 60.5$.

Section 2.2: Data Cleaning and Preprocessing Techniques

2.2.1 Common Data Cleaning Techniques

1. **Handling Missing Values**: Missing data can occur for various reasons, such as errors in data collection or transfer. There are several ways to handle missing data, depending on the nature and extent of the missing values.

   - **Imputation**: Here, you replace missing values with substituted values. The substituted value could be a central tendency measure like mean, median, or mode for numerical data or the most frequent category for categorical data. More sophisticated imputation methods include regression imputation and multiple imputation.
   - **Deletion**: You remove the instances with missing values from the dataset. While this method is straightforward, it can lead to loss of information, especially if the missing data is not random.

2. **Removing Duplicates**: Duplicate entries can occur for various reasons, such as data entry errors or data merging. These duplicates can skew the data and lead to biased results. Techniques for removing duplicates involve identifying these redundant entries based on key attributes and eliminating them from the dataset.

3. **Data Type Conversion**: Sometimes, the data may be in an inappropriate format for a particular analysis or model. For instance, a numerical attribute may be recorded as a string. In such cases, data type conversion, also known as *datacasting*, is used to change the data type of a particular attribute or set of attributes. This process involves

converting the data into a suitable format that machine learning algorithms can easily process.

4. **Outlier Detection**: Outliers are data points that significantly deviate from other observations. They can be caused by variability in the data or errors. Outlier detection techniques are used to identify these anomalies. These techniques include statistical methods, such as the Z-score or IQR method, and machine learning methods, such as clustering or anomaly detection algorithms.

2.2.2 Data Processing

Data preprocessing is critical in data science, particularly for machine learning applications. It involves preparing and cleaning the dataset to make it more suitable for machine learning algorithms. This process can reduce complexity, prevent overfitting, and improve the model's overall performance.

Common Data Preprocessing Techniques:

1. Data Scaling
    - Data scaling is a technique used to standardize the range of independent variables or features of data. It aims to standardize the data's range of features to prevent any feature from dominating the others, especially when dealing with large datasets. This is a crucial step in data preprocessing, particularly for algorithms sensitive to the range of the data, such as deep learning models.
2. Encoding Categorical Variables
    - Machine learning models require inputs to be numerical. If your data contains categorical data, you must encode them to numerical values before fitting and evaluating a model. This process, known as encoding categorical variables, is a common data preprocessing technique.
3. Data Splitting
    - Data Splitting is a technique to divide the dataset into two or three sets, typically training, validation, and test sets. You use the training set to train the model and the validation set to tune the model's parameters. The test set provides an unbiased evaluation of the final model. This technique is essential when dealing with large data, as it ensures the model is not overfitted to a particular subset of data.
4. Handling Missing Values
    - Missing data in the dataset can lead to misleading results. Therefore, it's essential to handle missing values appropriately. Techniques for handling missing values include deletion, removing the rows with missing values, and imputation, replacing the missing values with statistical measures like mean, median, or model. This step is crucial in ensuring the quality of data used for training machine learning models.

5. Feature Selection

    - Feature selection is a process in machine learning where you automatically select those features in your data that contribute most to the prediction variable or output in which you are interested. Having irrelevant features in your data can decrease the accuracy of many models, especially linear algorithms like linear and logistic regression. This process is particularly important for data scientists working with high-dimensional data, as it reduces overfitting, improves accuracy, and reduces training time.

2.2.3 Dealing with Missing Values and Outliers

Question: Why missing values treatment is required?

Missing data in the data set can reduce the power/fit of a model or can lead to a biased model because we have not analyzed the behavior and relationship with other variables correctly. It can lead to wrong prediction or classification.

Question: Why does my data have missing values?

We looked at the importance of treatment of missing values in a dataset. Following are the reasons for occurrence of these missing values. They may occur at two stages:

1. Data Extraction

    - It is possible that there are problems with extraction process. In such cases, we should double- check for correct data with data guardians. Some hashing procedures can also be used to make sure data extraction is correct. Errors at data extraction stage are typically easy to find and can be corrected easily as well.

2. Data collection

    - These errors occur at time of data collection and are harder to correct. They can be categorized in four types:

        (a) **Missing completely at random (MCAR)**. This is a case when the probability of missing variable is same for all observations. For example: respondents of data collection process decide that they will declare their earning after tossing a fair coin. If a head occurs, respondent declares his / her earnings and vice versa. Here each observation has equal chance of missing value.

        (b) **Missing at random (MAR)**. This is a case when variable is missing at random and missing ratio varies for different values / level of other input variables. For example: We are collecting data for age and female has higher missing value compare to male.

(c) **Missing that depends on unobserved predictors**. This is a case when the missing values are not random and are related to the unobserved input variable. For example: In a medical study, if a particular diagnostic causes discomfort, then there is higher chance of drop out from the study. This missing value is not at random unless we have included "discomfort" as an input variable for all patients.

(d) **Missing that depends on the missing value itself**. This is a case when the probability of missing value is directly correlated with missing value itself. For example: People with higher or lower income are likely to provide non-response to their earning.

2.2.4 Methods to Treat Missing Values

1. Deletion

- It is of two types: List Wise Deletion and Pair Wise Deletion.
  - In *listwise deletion*, we delete observations where any of the variable is missing. Simplicity is one of the major advantage of this method, but this method reduces the power of model because it reduces the sample size.
  - In *pairwise deletion*, we perform analysis with all cases in which the variables of interest are present. Advantage of this method is, it keeps as many cases available for analysis. One of the disadvantage of this method, it uses different sample size for different variables.
    Note: Deletion methods are used when the nature of missing data is "Missing completely at random (MCAR)" else non-random missing values can bias the model output.
    Example:

LISTWISE DELETION

Gender	Manpower	Sales
M	25	343
F	***	280
M	33	332
M	***	272
F	25	***
M	29	326
***	26	259
M	32	297

PAIRWISE DELETION

Gender	Manpower	Sales
M	25	343
F	***	280
M	33	332
M	***	272
F	25	***
M	29	326
***	26	259
M	32	297

2. Mean/ Mode/ Median Imputation

   - Imputation is a method to fill in the missing values with estimated ones. The objective is to employ known relationships that can be identified in the valid values of the data set to assist in estimating the missing values. Mean / Mode / Median imputation is one of the most frequently used methods. It consists of replacing the missing data for a given attribute by the mean or median (quantitative attribute) or mode (qualitative attribute) of all known values of that variable. It can be of two types:
     (a) **Generalized Imputation**. In this case, we calculate the mean or median for all non-missing values of that variable then replace missing value with mean or median. Like in above table, variable "Manpower" is missing so we take average of all non-missing values of "Manpower" (28.33) and then replace missing value with it.
     (b) **Similar case Imputation**. In this case, we calculate average for gender "Male" (29.75) and "Female" (25) individually of non-missing values then replace the missing value based on gender. For "Male", we will replace missing values of manpower with 29.75 and for "Female" with 25.

3. Prediction Model

   - Prediction model is one of the sophisticated method for handling missing data. Here, we create a predictive model to estimate values that will substitute the missing data. In this case, we divide our data set into two sets: One set with no missing values for the variable and another one with missing values. First data set become training data set of the model while second data set with missing values is test data set and variable with missing values is treated as target variable. Next, we create a model to predict target variable based on other attributes of the training data set and populate missing values of test data set. We can use regression, ANOVA, Logistic regression and various modeling technique to perform this.
   - There are 2 drawbacks for this approach:
     (a) The model estimated values are usually more well-behaved than the true values.
     (b) If there are no relationships with attributes in the data set and the attribute with missing values, then the model will not be precise for estimating missing values.

4. K-nearest Neighbor (KNN) Imputation

   - In this method of imputation, the missing values of an attribute are imputed using the given number of attributes that are most similar to the attribute whose values are missing. The similarity of two attributes is determined using a distance function. It is also known to have certain advantage and disadvantages.
   - *Advantages*:
     k-nearest neighbor can predict both qualitative and quantitative attributes. Creation of predictive model for each attribute with missing data is not required. Attributes with multiple missing values can be easily treated. Correlation structure of the data is taken into consideration.

- *Disadvantage*:
  KNN algorithm is very time-consuming in analyzing large database. It searches through all the dataset looking for the most similar instances. Choice of k-value is very critical. Higher value of k would include attributes which are significantly different from what we need whereas lower value of k implies missing out of significant attributes.

2.2.5 Techniques of Outlier Detection and Treatment

Outlier is a commonly used terminology by analysts and data scientists as it needs close attention else it can result in wildly wrong estimations. Simply speaking, Outlier is an observation that appears far away and diverges from an overall pattern in a sample.

1. Sorting Method
   You can sort quantitative variables from low to high and scan for extremely low or extremely high values. Flag any extreme values that you find. This is a simple way to check whether you need to investigate certain data points before using more sophisticated methods.
   Example:
   Your dataset for a pilot experiment consists of 8 values.

   $$180, 156, 9, 176, 163, 1827, 166, 171$$

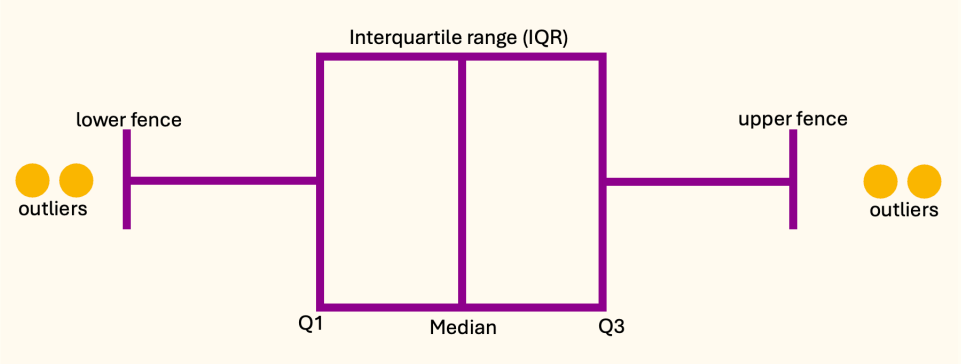
   You sort the values from low to high and scan for extreme values.

   $$\mathbf{9}, 156, 163, 166, 171, 176, 180, \mathbf{1872}$$

2. Using visualizations
   You can use software to visualize your data with a box plot, or a box-and-whisker plot, so you can see the data distribution at a glance. This type of chart highlights minimum and maximum values (the range), the median, and the interquartile range for your data.
   Many computer programs highlight an outlier on a chart with an asterisk, and these will lie outside the bounds of the graph.

3. Statistical outlier detection
   Statistical outlier detection involves applying statistical tests or procedures to identify extreme values. You can convert extreme data points into z scores that tell you how many standard deviations away they are from the mean.
   If a value has a high enough or low enough z score, it can be considered an outlier. As a rule of thumb, values with a z score greater than 3 or less than –3 are often determined to be outliers.

4. Using the interquartile range
   The interquartile range (IQR) tells you the range of the middle half of your dataset. You can use the IQR to create “fences” around your data and then define outliers as any values that fall outside those fences.



This method is helpful if you have a few values on the extreme ends of your dataset, but you aren't sure whether any of them might count as outliers.
Interquartile range method

(a) Sort your data from low to high
(b) Identify the first quartile ($Q1$), the median, and the third quartile ($Q3$).
(c) Calculate your $IQR = Q3–Q1$
(d) Calculate your upper fence = $Q3 + (1.5 * IQR)$
(e) Calculate your lower fence = $Q1–(1.5 * IQR)$
(f) Use your fences to highlight any outliers, all values that fall outside your fences.

Your outliers are any values greater than your upper fence or less than your lower fence.

2.2.6 Dealing with Outliers

Once you've identified outliers, you'll decide what to do with them. Your main options are retaining or removing them from your dataset. This is similar to the choice you're faced with when dealing with missing data.

For each outlier, think about whether it's a true value or an error before deciding.

- Does the outlier line up with other measurements taken from the same participant?
- Is this data point completely impossible or can it reasonably come from your population?
- What's the most likely source of the outlier? Is it a natural variation or an error?

In general, you should try to accept outliers as much as possible unless it's clear that they represent errors or bad data.

2.2.7 Retain Outliers

Just like with missing values, the most conservative option is to keep outliers in your dataset. Keeping outliers is usually the better option when you're not sure if they are errors.

With a large sample, outliers are expected and more likely to occur. But each outlier has less of an effect on your results when your sample is large enough. The central tendency and variability of your data won't be as affected by a couple of extreme values when you have a large number of values.

If you have a small dataset, you may also want to retain as much data as possible to make sure you have enough statistical power. If your dataset ends up containing many outliers, you may need to use a statistical test that's more robust to them. Non-parametric statistical tests perform better for these data.

2.2.8 Remove Outliers

Outlier removal means deleting extreme values from your dataset before you perform statistical analyses. You aim to delete any dirty data while retaining true extreme values.

It's a tricky procedure because it's often impossible to tell the two types apart for sure. Deleting true outliers may lead to a biased dataset and an inaccurate conclusion.

For this reason, you should only remove outliers if you have legitimate reasons for doing so. It's important to document each outlier you remove and your reasons so that other researchers can follow your procedures.

Chapter 3
Descriptive Statistics

Section 3.1: Frequencies and Descriptive Statistics

Effective presentation of study results, in presentation or manuscript form, typically starts with frequencies and descriptive statistics (ie, mean, medians, standard deviations).

One can get a better sense of the variables by examining these data to determine whether a balanced and sufficient research design exists. Frequencies also inform on missing data and give a sense of outliers.

Luckily, software programs are available to conduct exploratory data analysis.

Example:
Research Question: Are there differences in drug life (length of effect) for Drug 23 based on the administration site?
A more precise hypothesis could be: "Is drug 23 longer-lasting when administered via site A compared to site B?"

First, it is essential to start with the frequencies of the variables. To keep things simple, only variables of minutes (drug life effect) and administration site (A vs B) are included.

Administration Site

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	A	50	50.0	50.0	50.0
	B	50	50.0	50.0	100.0
	Total	100	100.0	100.0	

Minutes

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	135.00	1	1.0	1.0	1.0
	306.00	1	1.0	1.0	2.0
	481.50	1	1.0	1.0	3.0
	486.00	2	2.0	2.0	5.0
	490.50	2	2.0	2.0	7.0
	495.00	3	3.0	3.0	10.0
	499.50	7	7.0	7.0	17.0
	504.00	4	4.0	4.0	21.0

The figure shows that the administration site appears to be a balanced design with 50 individuals in each group. The excerpt for minutes frequencies is the bottom portion of the figure and shows how many cases fell into each time frame with the cumulative percent on the right-hand side.

In examining the figure, one suspiciously low measurement (135) was observed, considering time variables. If a data point seems inaccurate, a researcher should find this case and confirm if this was an entry error.

For the sake of this review, the authors state that this was an entry error and should have been entered 535 and not 135. Had the analysis occurred without checking this, the data analysis, results, and conclusions would have been invalid.

When finding any entry errors and determining how groups are balanced, potential missing data is explored. If not responsibly evaluated, missing values can nullify results.

After replacing the incorrect 135 with 535, descriptive statistics, including the mean, median, mode, minimum/maximum scores, and standard deviation were examined. Output for the research example for the variable of minutes can be seen in the figure on the right.

Statistics

Minutes

N	Valid	100
	Missing	0
Mean		535.0585
Median		535.0000
Mode		499.50
Std. Deviation		50.15193
Variance		2515.216
Minimum		306.00
Maximum		895.00

Observe each variable to ensure that the mean seems reasonable and that the minimum and maximum are within an appropriate range based on medical competence or an available codebook. One assumption common in statistical analyses is a normal distribution.

The figure shows that the mode differs from the mean and the median. We have visualization tools such as histograms to examine these scores for normality and outliers before making decisions.

3.1.1 Histograms

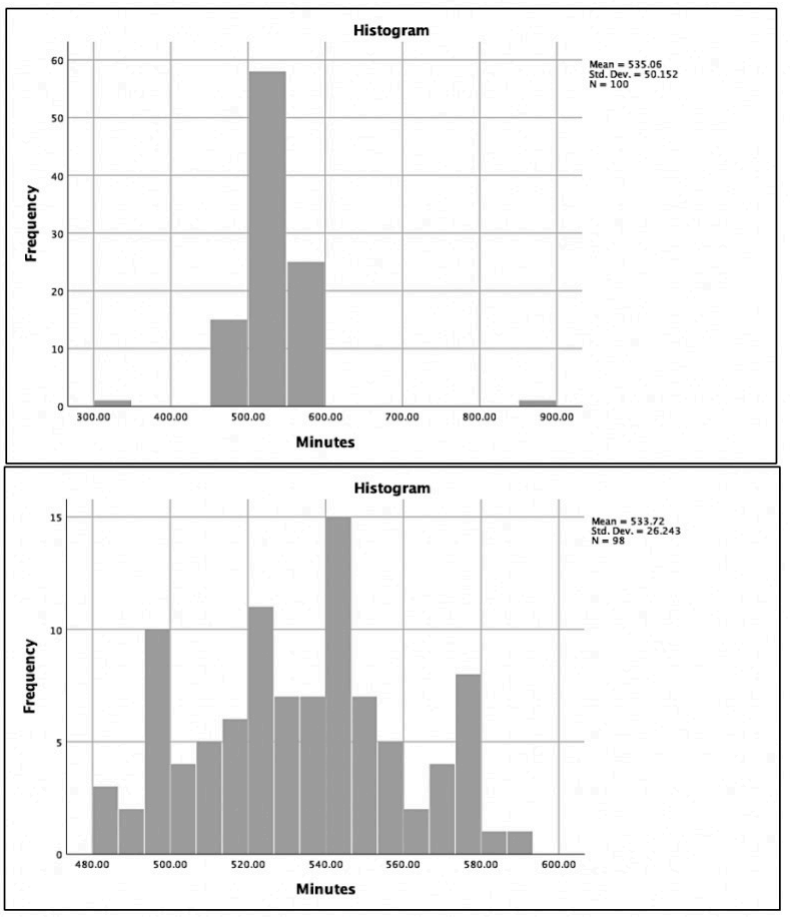
Histograms are useful in assessing normality, as many statistical tests (eg, ANOVA and regression) assume the data have a normal distribution. When data deviate from a normal distribution, it is quantified using skewness and kurtosis.

Skewness occurs when one tail of the curve is longer. If the tail is lengthier on the left side of the curve (more cases on the higher values), this would be negatively skewed, whereas if the tail is longer on the right side, it would be positively skewed.

Kurtosis is another facet of normality. Positive kurtosis occurs when the center has many values falling in the middle, whereas negative kurtosis occurs when there are very heavy tails.

Additionally, histograms reveal outliers: data points either entered incorrectly or truly very different from the rest of the sample.

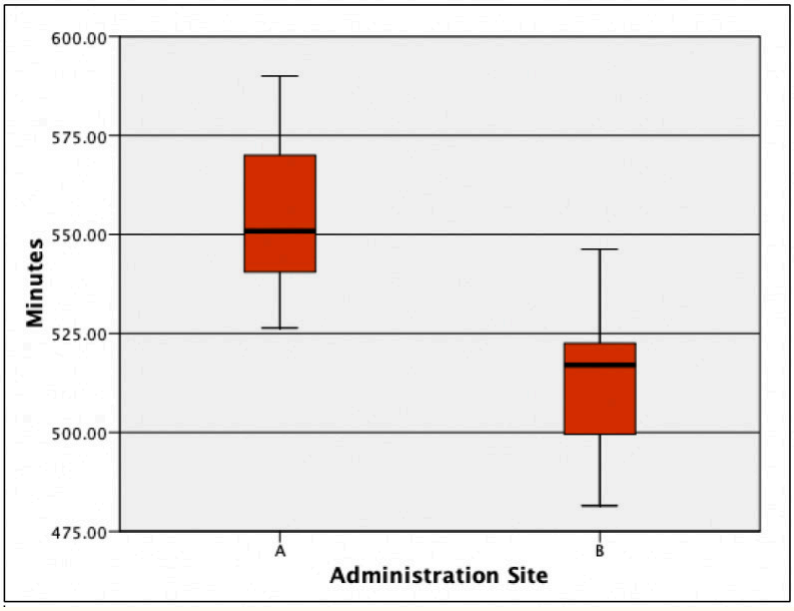
The figure below provides an example of a histogram. In the figure, 2 possible outliers causing kurtosis are observed. This histogram appears much closer to an approximately normal distribution with the kurtosis being treated. Remember, all evidence should be considered before eliminating outliers. When reporting outliers in scientific paper outputs, account for the number of outliers excluded and justify why they were excluded.



3.1.2 Boxplots

Boxplots can examine for outliers, assess the range of data, and show differences among groups. Boxplots provide a picture of data distribution when there are numerous values, and all values cannot be displayed.

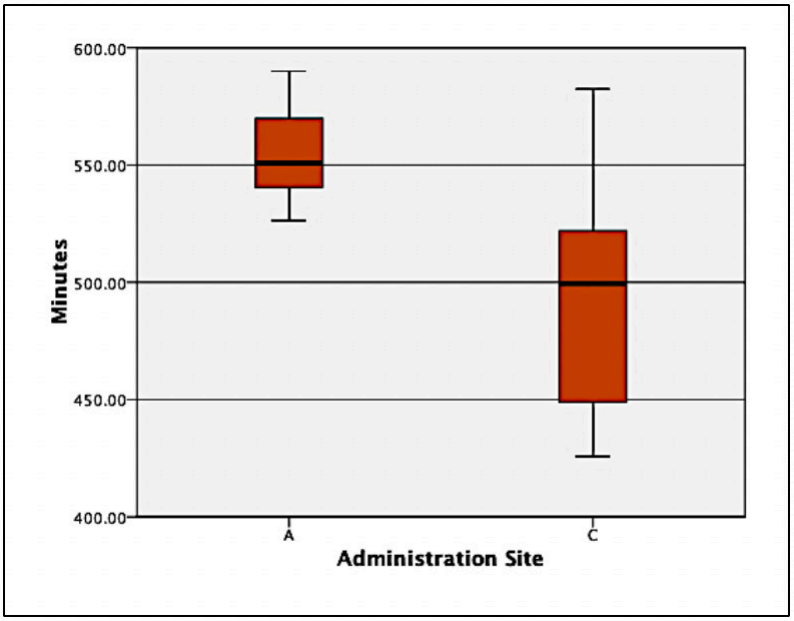
The figure below show differences with potential clinical impact. Had any outliers existed, they would appear outside the line endpoint. The red boxes represent the middle 50% of scores. The lines within each red box represent the median number of minutes within each administration site. The horizontal lines at the top and bottom of each line connected to the red box represent the 25^{th} and 75^{th} percentiles.



In examining the difference boxplots, an overlap in minutes between 2 administration sites were observed: the approximate top 25 percent from site B had the same time noted as the bottom 25 percent at site A.

Site B had a median minute amount under 525, whereas administration site A had a length greater than 550. If there were no differences in adverse reactions at site A, analysis of this figure provides evidence that healthcare providers should administer the drug via site A. Researchers could follow by testing a third administration site, site C.

Figure below displays the same site A data as Figure 4, but something looks different. The significant variance at site C makes site A's variance appear smaller.



In order words, patients who were administered the drug via site C had a larger range of scores. Thus, some patients experience a longer half-life when the drug is administered via site C than the median of site A; however, the broad range (lack of accuracy) and lower median should be the focus.

The precision of minutes is much more compacted in site A. Therefore, the median is higher, and the range is more precise. One may conclude that this makes site A a more desirable site.

3.1.3 Significance of Descriptive Statistics

Ultimately, by understanding basic exploratory data methods, researchers and consumers of research can make quality and data-informed decisions. These data-informed decisions will result in the ability to appraise the significance of research outputs. By overlooking these fundamentals in statistics, critical errors in judgment can occur.

Chapter 4
Data Visualization

Section 4.1: Presentation of Data

Different types of data presentation

1. Textual presentation
2. Tabular presentation
3. Graphical presentation

4.1.1 Textual Presentation

Textual presentation of data incorporates important figures in a paragraph of text. In textual presentation, we insert important data figures or summary measures within the paragraph of text to support our conclusions and answers to the research problem. At the same time, we can also highlight all the noteworthy figures in understanding the population we are studying.

Example:

> "At last count, 38 airlines were operating Boeing 707's, 720's, and 727's over the world's airlines. The far-flung Boeing fleet has now logged an estimated 1,803,704,000 miles (22,855,948,000 kms) and has massed approximately 4,096,000 revenue flight hours. Passenger totals stand at upwards of 71.6 million."

4.1.2 Tabular Presentation

Tabular presentation of data arranges figures in a systematic manner in rows and columns. Being the most common method of data presentation, we can use it for various purposes

such as description, comparison, and even showing relationships between two or more variables of interest.

Types of Tabular Presentation
There are three different types of tabular presentation which vary in their format and layout.

1. Leader Work
   Leader work has the simplest layout among all three types of tables. It contains no table title or column headings and has no table borders.
   Example:

The population in the Philippines for the census years 1975 to 2000 is as follows:

1975	42,070,660
1980	48,098,460
1990	60,703,206
1995	68,616,536
2000	76,498,735

Source: National Statistics Office

2. Text Tabulation
   The format of text tabulation is a little bit more complex than leader work. It already has column headings and table borders, making it easier to understand than leader work. However, just like it, it still has no table title and table number.
   Example:

The distribution of cellular subscribers per telephone as of December 2003 is as follows:

Telephone Operator	Number of Subscribers
Smart	10,080,112
Globe Telecom	8,800,000
Piltel	2,867,085
Extelcom	29,896
TOTAL	22,509,560

Source: National Telecommunication Commission

3. Formal Statistical Table
   The formal statistical table is the most complete type of table since it has all the different and essential parts of a table like table number, table title, head note, box head,

stub head, column headings, and so on. It is a stand-alone table. Even without a descriptive text or introductory statement, the reader should easily understand its contents.
Example:

Table 4.4 - CRIME VOLUME AND RATE BY TYPE: 1991-1993
(Rate per 100,000 population)

Type	1991		1992		1993	
	Volume	Crime Rate	Volume	Crime Rate	Volume	Crime Rate
Total	121,236	195	104,719	164	96,686	148
Index Crimes	77,261	124	1 67,354	106	58,684	90
Murder	8,707	14	8,293	13	7,758	12
Homicide	8,069	13	7,912	12	7,123	11
Physical Injury	21,862	35	20,462	32	18,722	29
Robbery	13,817	22	11,164	18	9,856	15
Theft	22,780	37	17,372	27	12,940	20
Rape	2,026	3	2,149	3	2,285	4
Nonindex Crimes	44,065	71	37,365	59	38,002	58

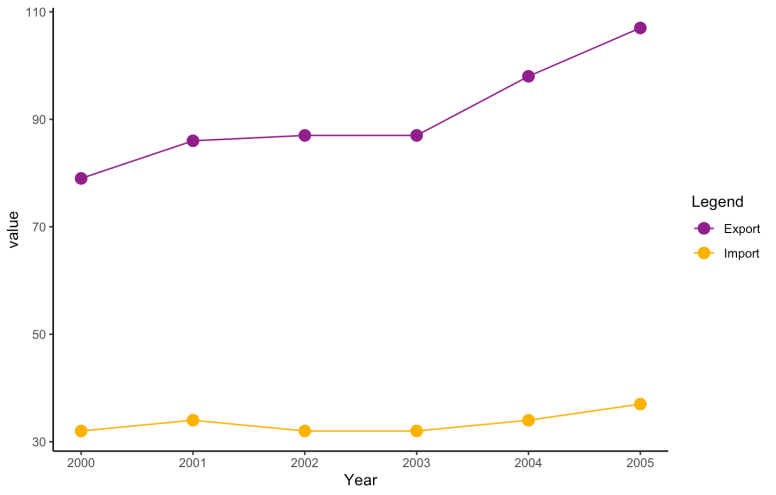
Source: Philippine National Police

4.1.3 Graphical Presentation

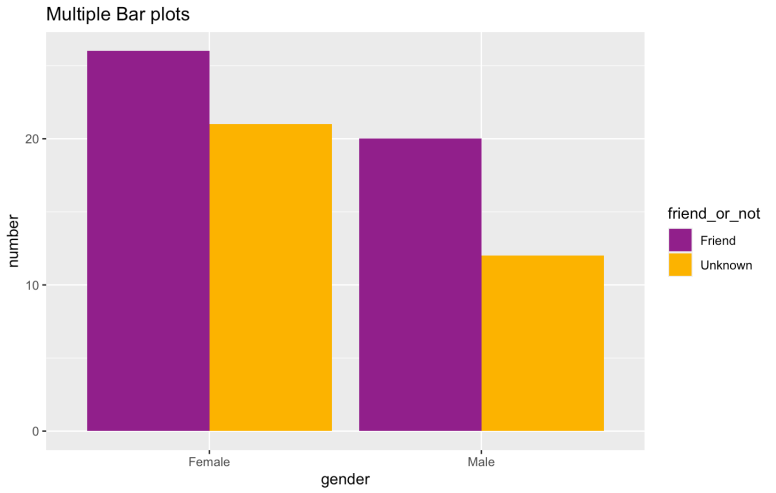
Graphical presentation of data portrays numerical figures or relationships among variables in pictorial form. In constructing a good chart, it has to be accurate, clear, simple, professional, and well designed.

The different types of statistical charts are the following:

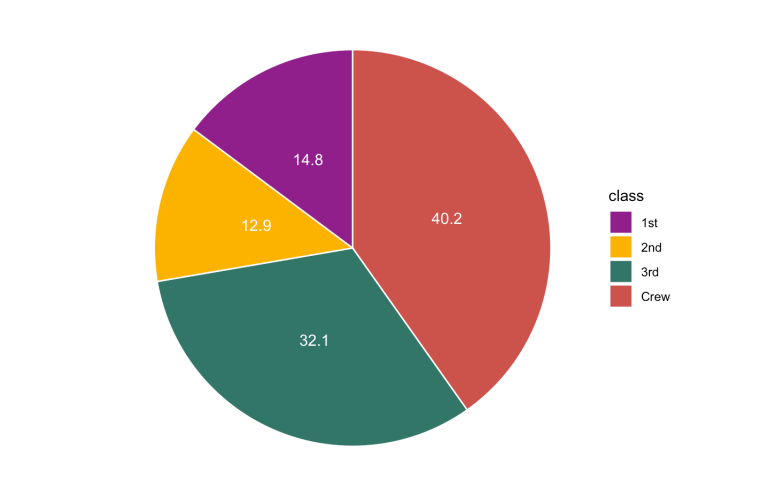
1. Line chart
   The line chart is useful for presenting historical data. This chart is effective in showing the movement of a series over time. The movement may be increasing, decreasing, stationary/constant, or fluctuating.
   Example:



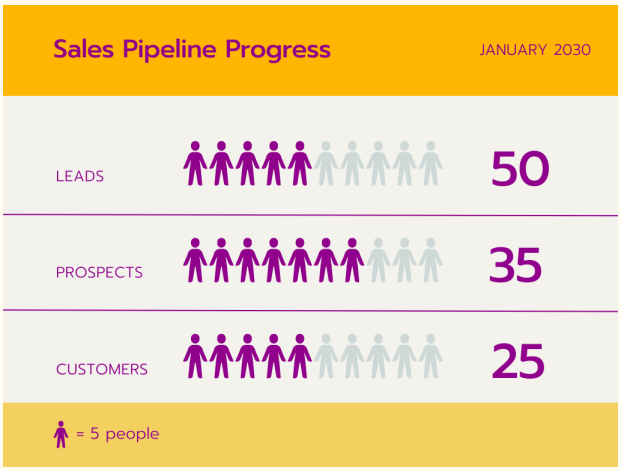
2. Bar chart or Column chart
   The column/bar chart is useful for comparing amounts in a time series data. The emphasis in a column chart is on the magnitude rather than the movement of a series.
   Example:



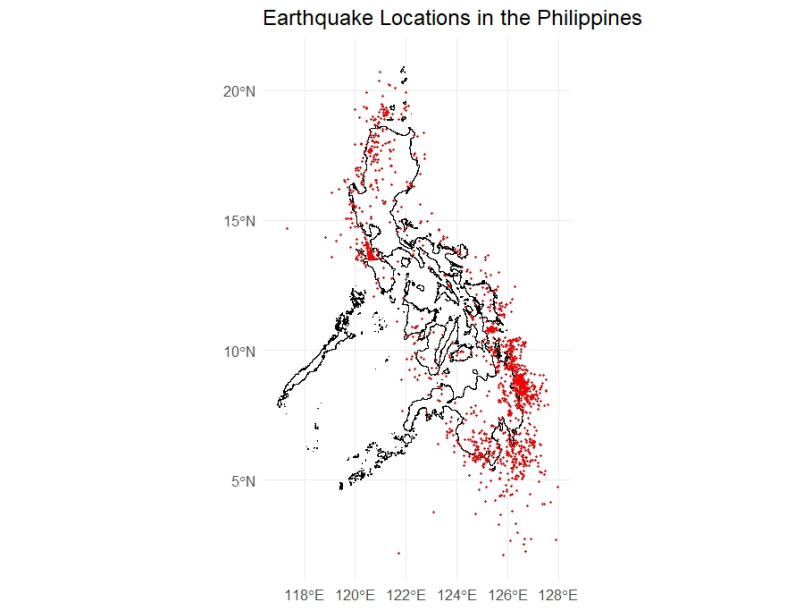
3. Pie chart
   The pie chart is useful for data sorted into categories for a specific period. The purpose is to show the component parts with respect to the total in terms of the percentage distribution. Each section or slice indicates the proportion of each component or category. It is applicable for qualitative rather than quantitative data. We use the pie chart if the variable has less than 6 categories.
   Example:



4. Pictogram
   The pictograph is useful to get the attention of the reader while providing an overall picture of the data without presenting the exact figures. However, it still allows the comparison of different categories even if we present the approximate values only.
   Example:



5. Statistical maps
   The pictograph is useful in showing statistical data in geographical areas. We also call these types of charts as crosshatched maps or shaded maps. Geographic areas may be barangays, cities, districts, provinces, or countries. The figures in the map can be ratios, rates, percentages, or indices.
   Example:


Earthquake Locations in the Philippines
20°N
15°N
10°N
5°N
118°E
120°E
122°E
124°E
126°E
128°E

Chapter 5
EDA Techniques

Section 5.1: Correlation and Covariance

Covariance and correlation are two mathematical concepts used in statistics. Both terms are used to describe how two variables relate to each other. Covariance is a measure of how two variables change together.

The terms covariance vs correlation is very similar to each other in probability theory and statistics. Both terms describe the extent to which a random variable or a set of random variables can deviate from the expected value.

Covariance can be positive, negative, or zero. A positive covariance means that the two variables tend to increase or decrease together. A negative covariance means that the two variables tend to move in opposite directions.

A zero covariance means that the two variables are not related. Correlation can only be between -1 and 1. A correlation of -1 means that the two variables are perfectly negatively correlated, which means that as one variable increases, the other decreases.

A correlation of 1 means that the two variables are perfectly positively correlated, which means that as one variable increases, the other also increases. A correlation of 0 means that the two variables are not related.

5.1.1 What is Covariance?

Covariance signifies the direction of the linear relationship between the two variables. By direction we mean if the *variables* are directly proportional or inversely proportional to each other. (Increasing the value of one variable might have a positive or a negative impact on the value of the other variable).

The values of covariance can be any number between the two opposite infinities. Also, it's important to mention that covariance only measures how two variables change together, not the dependency of one variable on another one.

The value of covariance between 2 variables is achieved by taking the summation of the product of the differences from the means of the variables as follows:

$$Cov(x, y) = \frac{\sum(X_i - \bar{X}) \sum(Y_i - \bar{Y})}{N}$$

covariance is only useful to find the direction of the relationship between two variables and not the magnitude. Below are the plots which help us understand how the covariance between two variables would look in different directions.

5.1.2 What is Correlation?

Correlation analysis is a method of statistical evaluation used to study the strength of a relationship between two, numerically measured, continuous variables.

It not only shows the kind of relation (in terms of direction) but also how strong the relationship is. Thus, we can say the correlation values have standardized notions, whereas the covariance values are not standardized and cannot be used to compare how strong or weak the relationship is because the magnitude has no direct significance. It can assume values from -1 to +1.

To determine whether the covariance of the two variables is large or small, we need to assess it relative to the standard deviations of the two variables.

To do so we have to normalize the covariance by dividing it with the product of the standard deviations of the two variables, thus providing a correlation between the two variables.

The main result of a correlation is called the correlation coefficient.

$$Correlation = \frac{Cov(x, y)}{\sigma_x \sigma_y}$$

where Cov is the covariance, σ_x and σ_y are the standard deviations of X and Y, respectively.

5.1.3 Difference Between Correlation and Covariance

Aspect	Covariance	Correlation
Definition	Measures the joint variability of two random variables.	Measures the strength and direction of the linear relationship between two variables.
Range	Can take any value from negative infinity to positive infinity	Ranges from -1 to 1.
Units	Has units – the product of the units of the two variables.	Dimensionless (no units), a standardized measure.
Normalization	Not normalized – the magnitude depends on the units of the variables.	Normalized – independent of the scale of variables.
Interpretation	Difficult to interpret the strength of the relationship due to lack of normalization.	Easy to interpret because it's a standardized coefficient (usually Pearson's r).
Sensitivity	Sensitive to the scale and units of measurement of the variables.	Not sensitive to the scale and units of measurement since it's a relative measure.

Section 5.2: Re-expressing Data

Datasets do not always follow a nice symmetrical distribution nor do their spreads behave systematically across different levels (e.g. medians). Such distributions do not lend themselves well to visual exploration since they can mask simple patterns. They can also be a problem when testing hypotheses using traditional statistical procedures.

A solution to this problem is non-linear **re-expression** (aka transformation) of the values. In univariate analysis, we often seek to **symmetrize** the distribution and/or **equalize** the spread. In multivariate analysis, the objective is to usually **linearize** the relationship between variables and/or to **normalize** the residual in a regression model.

Re-expressing values consist of changing the scale of measurement from what is usually a linear scale to a non-linear scale. One popular form of re-expression is the `log`.

5.2.1 The log Transformation

One of the most popular transformations used in data analysis is the *logarithm*. The log, $\log_b(x)$, of a value is the power to which the base must be raised to produce . This requires that the log function be defined by a **base**, b, such as 10, 2 or exp(1) (the latter defining the natural log).

$$y = \log_b x \iff x = b^y$$

A log transformation preserves the order of the observations as measured on a linear scale but modifies the "distance" between them in a systematic way.

The log is not the only transformation that can be applied to a dataset. There is a whole family of power transformations (of which the log is a special case) that can be implemented using either the *Tukey transformation* or the *Box-Cox transformation*.

5.2.2 The Tukey Transformation

The Tukey family of transformations offers a broader range of re-expression options (which includes the log). The values are re-expressed using the algorithm:

$$T_{Tukey} = \begin{cases} x^p, & p \neq 0 \\ \log(x), & p = 0 \end{cases}$$

The objective is to find a value for from a "ladder" of powers (e.g. $-2, -1, -1/2, 0, 1/2, 1, 2$) that does a good job in re-expressing the batch of values. Technically, can take on any value. But in practice, we normally pick a value for that may be "interpretable" in the context of our analysis. For example, a log transformation ($p = 0$) may make sense if the process we are studying has a steady growth rate. A cube root transformation ($p = 1/3$) may make sense if the entity being measured is a volume (e.g. rain fall measurements).

But sometimes, the choice of may not be directly interpretable or may not be of concern to the analyst. A nifty solution to finding an appropriate is to create a function whose input is the vector (that we want to re-express) and a parameter we want to explore.

5.2.3 Box-Cox Transformation

Another family of transformations is the Box-Cox transformation. The values are re-expressed using a modified version of the Tukey transformation:

$$T_{Box-Cox} = \begin{cases} \frac{x^p - 1}{p}, & p \neq 0 \\ \log(x), & p = 0 \end{cases}$$

The choice of re-expression will depend on the analysis context. For example, if you want an easily interpretable transformation then opt for the Tukey re-expression. If you want to compare the shape of transformed variables, the Box-Cox approach will be better suited.

Section 5.3: Median Polish

Median Polish is a robust statistical technique used in exploratory data analysis (EDA) to decompose a two-way table or a multi-way array into components, thereby simplifying the analysis of complex data structures. It's particularly useful when dealing with large datasets or datasets with outliers and missing values.

5.3.1 Data Structure

Median Polish is typically applied to a two-way table, often representing data collected in a factorial experiment or observational study. However, it can also be extended to multi-way arrays in higher dimensions.

5.3.2 Decomposition

The main idea behind Median Polish is to iteratively subtract row and column medians from the data until convergence. This process effectively decomposes the original table into three components: row effects, column effects, and residuals.

The row effects capture the systematic variation associated with each row, the column effects capture the systematic variation associated with each column, and the residuals represent the remaining variation not accounted for by the row and column effects.

5.3.3 Robustness

One of the key strengths of Median Polish is its robustness to outliers and skewness in the data. By using medians rather than means, Median Polish is less sensitive to extreme values, making it suitable for datasets with non-normal distributions or data points with high variability.

This robustness makes Median Polish particularly valuable in exploratory analysis of real-world datasets where outliers or non-normality are common.

5.3.4 Interpretation

After applying Median Polish, the decomposed components can be interpreted to understand the underlying structure of the data. For example, significant row effects may indicate differences between groups or conditions, while significant column effects may suggest patterns or trends across different variables or factors.

5.3.5 Applications

Median Polish is widely used in various fields such as environmental science, agriculture, biology, and social sciences for exploratory analysis of complex datasets.

It can help researchers identify interesting patterns, relationships, or anomalies in the data before conducting more formal statistical analyses or hypothesis testing.

Chapter 6
References

1. Anscombe, F. (1973), Graphs in Statistical Analysis, The American Statistician, pp. 195-199.
2. Anscombe, F. and Tukey, J. W. (1963), The Examination and Analysis of Residuals, Technometrics, pp. 141-160.
3. Barnett and Lewis (1994), Outliers in Statistical Data, 3rd. Ed., John Wiley and Sons.
4. Bhandari, P. (2024, January 17). How to Find Outliers | 4 Ways with Examples & Explanation. Scribbr. Retrieved February 29, 2024, from https://www.scribbr.com/statistics/outliers/
5. Birnbaum, Z. W. and Saunders, S. C. (1958), A Statistical Model for Life-Length of Materials, Journal of the American Statistical Association, 53(281), pp. 151-160.
6. Bloomfield, Peter (1976), Fourier Analysis of Time Series, John Wiley and Sons.
7. Box, G. E. P. and Cox, D. R. (1964), An Analysis of Transformations, Journal of the Royal Statistical Society, pp. 211-243, discussion pp. 244- 252.
8. Box, G. E. P., Hunter, W. G., and Hunter, J. S. (1978), Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building, John Wiley and Sons.
9. Box, G. E. P., and Jenkins, G. (1976), Time Series Analysis: Forecasting and Control, Holden-Day.
10. Bradley, (1968). Distribution-Free Statistical Tests, Chapter
11. Brase, C. H., & Brase, C. P. (2018). Understandable statistics: Concepts and methods (12th ed.). Cengage Learning.
12. Brown, M. B. and Forsythe, A. B. (1974), Journal of the American Statistical Association, 69, pp. 364-367.
13. Buttarazzi D, Pandolfo G, Porzio GC. A boxplot for circular data. Biometrics. 2018 Dec;74(4):1492-1501.
14. Chakravarti, Laha, and Roy, (1967). Handbook of Methods of Applied Statistics, Volume I, John Wiley and Sons, pp. 392-394.

15. Chambers, John, William Cleveland, Beat Kleiner, and Paul Tukey, (1983), Graphical Methods for Data Analysis, Wadsworth.

16. Chang, W., Cheng, J., Allaire, J. J., Xie, Y., & McPherson, J. (2021). shiny: Web Application Framework for R. R package version 1.7.1. Retrieved from https://CRAN.R-project.org/package=shiny

17. Cleveland, W. S. (1993). Visualizing data. Hobart Press.

18. Clin Chim Acta. 2006 Apr;366(1-2):112-29.

19. Few, S. (2009). Now you see it: Simple visualization techniques for quantitative analysis. Analytics Press.

20. Folch-Fortuny, A., Villaverde, A., Ferrer, A., & Banga, J. (2015). Enabling network inference methods to handle missing data and outliers. BMC Bioinformatics, 16(1). https://doi.org/10.1186/s12859-015-0717-7

21. Hazra A, Gogtay N. Biostatistics Series Module 1: Basics of Biostatistics. Indian J Dermatol. 2016 Jan- Feb;61(1):10-20.

22. Heart. 2016 Mar;102(5):349-55.

23. Henderson AR. Testing experimental data for univariate normality.

24. Hildebrand, D. K., & Ottoboni, A. (2019). Statistical thinking for managers. Routledge.

25. Kim, Hae-Young (2013-02-01). "Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis". Restorative Dentistry & Endodontics. 38 (1): 52–54. doi:10.5395/rde.2013.38.1.52. ISSN 2234-7658. PMC 3591587. PMID 23495371.

26. Kumar, N., Hoque, M., Shahjaman, M., Islam, S., & Mollah, M. (2017). Metabolomic biomarker identification in presence of outliers and missing values. Biomed Research International, 2017, 1-11. https://doi.org/10.1155/2017/2437608

27. Mowbray FI, Fox-Wasylyshyn SM, El-Masri MM. Univariate Outliers: A Conceptual Overview for the Nurse Researcher. Can J Nurs Res. 2019 Mar;51(1):31-37.

28. "New View of Statistics: Non-parametric Models: Rank Transformation". www.sportsci.org. Retrieved 2019-03-23.

29. Nugroho, H., Utama, N., & Surendro, K. (2021). Normalization and outlier removal in class center-based firefly algorithm for missing value imputation. Journal of Big data, 8(1). https://doi.org/10.1186/s40537-021-00518-7

30. Rice K, Lumley T. Graphics and statistics for cardiology: comparing categorical and continuous variables.

31. Sheng Y, Ge Y, Yuan L, Li T, Yin FF, Wu QJ. Outlier identification in radiation therapy knowledge-based planning: A study of pelvic cases. Med Phys. 2017 Nov;44(11):5617-5626.

32. ”Testing normality including skewness and kurtosis”. imaging.mrc-cbu.cam.ac.uk. Retrieved 2019-03-18.

33. Van Droogenbroeck F.J., ’An essential rephrasing of the Zipf-Mandelbrot law to solve authorship attribution applications by Gaussian statistics’ (2019) [1]

34. Warton, D.; Hui, F. (2011). ”The arcsine is asinine: the analysis of proportions in ecology”. Ecology. 92 (1): 3–10. doi:10.1890/10-0340.1. hdl:1885/152287. PMID 21560670.

35. Weissman C. Analyzing intensive care unit length of stay data: problems and possible solutions. Crit Care Med. 1997 Sep;25(9):1594-600.

36. Weisstein, Eric W. ”Covariance”. MathWorld.

37. Weisstein, Eric W. ”Statistical Correlation”. MathWorld.

38. Wickham, H. (2016). ggplot2: Elegant graphics for data analysis. Springer.

39. Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., & Woo, K. (2021). ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics. R package version 3.3.5. Retrieved from https://CRAN.R-project.org/package=ggplot2

40. Yuan, K., Marshall, L., & Bentler, P. (2002). A unified approach to exploratory factor analysis with missing data, nonnormal data, and in the presence of outliers. Psychometrika, 67(1), 95-121. https://doi.org/10.1007/bf02294711