

Exploratory Data Analysis

Descriptive Statistics & Outliers

Table of Contents

Part I: Introduction and Univariate Analysis

- What's EDA
- Key Concepts
- Data types
- EDA Framework
- Univariate Analysis techniques

Part II: Bivariate Analysis and Outliers

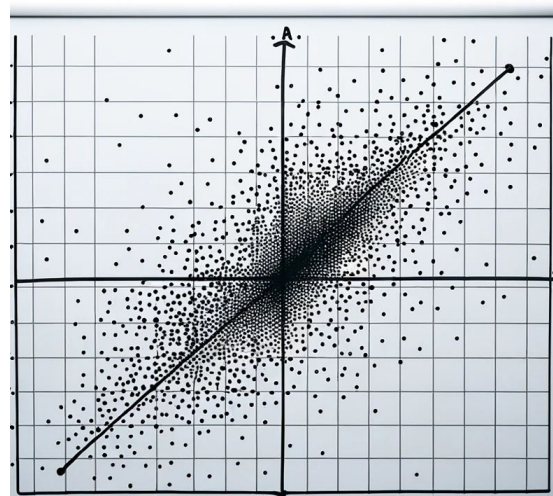
- Bivariate Analysis techniques
- Atypical values (Outliers)

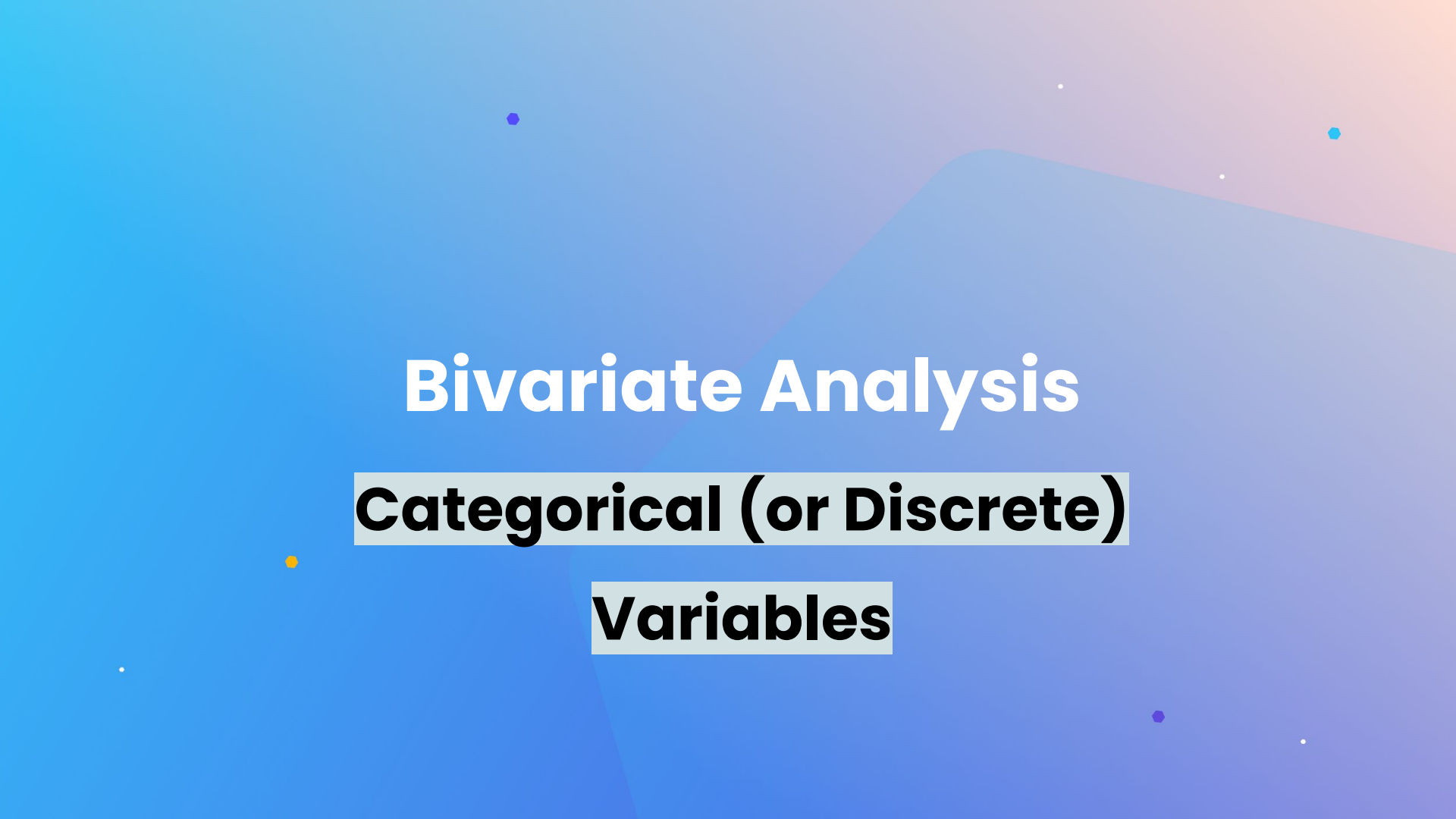
Bivariate Analysis

Exploring The Relationship and Interactions
Between Two Variables in Depth

Bivariate Analysis

- When working with **two categorical variables**, we focus on understanding the **relationships** and interactions between the **categories**.
- When working with **two numerical continuous variables**, we aim to determine the **strength** and **nature** of the **relationship** between them.
- When analyzing **one categorical and one numerical variable** together, we often look at how the **numerical** variable's **central tendency and spread** might **differ** across the different **categories**.





Bivariate Analysis

Categorical (or Discrete)

Variables

Two Categorical or Discrete Variables

Contingency Tables or Cross-tabulation

Table that represents the **frequency** distribution of categories from both variables. Used to understand the **relationship** between **two or more categorical variables**.

- Computes probabilities of the joint occurrence of two categorical variables.
- It is often used in hypothesis testing to determine if the observed frequencies for categories differ from the expected frequencies, e.g., Chi-square tests.

Index	Gender	Drink
0	Male	Tea
1	Male	Coffee
2	Male	Tea
3	Male	Coffee
4	Male	Coffee
5	Female	Tea
6	Female	Tea
7	Female	Tea
8	Female	Tea
9	Female	Coffee



	Coffee	Tea	Total
Female	1	4	5
Male	3	2	5
Total	4	6	10

- **Rows:** categories of one variable.
- **Columns:** categories of another variable.
- **Cells:** the count for a specific category combination.
- **Marginal frequencies:** row and column totals. Total counts for one variable disregarding the other variable.

Two Categorical or Discrete Variables

Chi-square Test

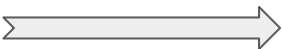
$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

χ^2 = the test statistic \sum = the sum of
O = Observed frequencies E = Expected frequencies

Determines if there's a **statistically significant association** between the two categorical variables.
It **compares**:

- **Observed Frequencies in a Contingency Table:** The actual counts in each cell of the table.
- **Expected Frequencies:** What we would expect the counts in each cell to be if there were no relationship between the two variables (**if the two variables were independent**).

	Coffee	Tea	Total
Female	1	4	5
Male	3	2	5
Total	4	6	10


$$\text{Expected frequency} = \frac{\text{Row Total} \times \text{Column Total}}{\text{Overall Total}}$$

	Coffee (Expected)	Tea (Expected)	Total
Female	2.0	3.0	5.0
Male	2.0	3.0	5.0
Total	4.0	6.0	10.0

Two Categorical or Discrete Variables

Chi-square Test

Note: We'll delve deeper into the Chi-Square Test later; now, we're emphasizing its application and result interpretation. We'll use Python, especially the *scipy* library, for the Chi-Square Test, bypassing manual calculations.

The key output is the **p-value**: a measure that helps us determine if the observed data deviates from what we would expect under certain assumptions—in this case, that the **two categorical variables are independent** of each other.

- **p-value < 0.05**: suggests that the observed data is significantly different from what we'd expect if the variables were independent. **Sufficient evidence to conclude there is a relationship between variables.**
- **p-value ≥ 0.05**: suggests that the observed data doesn't deviate much from what we'd expect under the assumption of independence. **No sufficient evidence to conclude that variables are related.**

Note: The p-value signals the existence of an association, not its strength or direction.

Assumptions:

- Assumes observations are independent.
- Requires a sufficiently large sample size. Cells should have expected frequencies of 5 or more.

Two Categorical or Discrete Variables

Chi-square Test: Example

A mobile store surveys 150 customers to discern smartphone brand preferences across three age groups: "Youth" (18-30), "Middle-Aged" (31-50), and "Seniors" (51+), aiming to refine its marketing strategies.

Observed Data (from survey):

	Apple	Samsung	Google
Youth	40	25	5
Middle-Aged	15	30	5
Seniors	10	20	10

Result: using `chi2_contingency()`, **p = 0.1487**

The p-value > 0.05 indicates no strong evidence of a significant association between age groups and smartphone brand preference based on our sample; the observed preferences might be coincidental. If $p < 0.05$, it'd imply a significant association, guiding the store's marketing strategies.

Hypothesis testing and p-value

Hypothesis testing is a statistical method which is used to make decision about entire population, with the help of only sample data. We will have two competing and non-overlapping hypothesis:

- **Null Hypothesis (H_0)**

- **Alternative Hypothesis (H_1 or H_a):** This is the argument which we would like to prove to be true.

This means we want to reject H_0 , to accept H_a .

The significance level is a probability of rejecting the null hypothesis when it is actually true. It is denoted by α and usually is **0.05**.

*To make this decision, we come up with a value called as **p-value**... the **p-value is a probability** of observing the results of the Null Hypothesis. So we want a very small P-value (smaller than α).*

To use a P-value to make a conclusion in a hypothesis test, compare P-value with α (0.05).

- If $P \leq \alpha$, then reject H_0

- If $P > \alpha$, then fail to reject H_0

If $p \leq \alpha$, there is less than 5% chance that the data being tested could have occurred if H_0 were true.

Two Categorical or Discrete Variables

Cramér's V – effect size

Cramér's V is a statistic used to **quantify the strength of association between two categorical variables**. It's an **effect size measurement** for the chi-square test of independence.

- 0: No association between the variables.
- 1: Perfect association (one variable perfectly predicts the other).
- There are multiple ways to interpret Cramér's V, relative to discipline and expectations of the experiment. One of them is from Cohen (1988) where the interpretation depends on the degrees of freedom:

df	small	medium	large
1	0.1	0.3	0.5
2	0.07	0.21	0.35
3	0.06	0.17	0.29

df here is the minimum number of categories in either rows or columns minus one.

- *For a 2x2 table (df = 1), for a 2x3 table (df = 1), for a 3x3 table (df = 2) and so on...*

In that chi2 **example**, if we run `association(observed_data, method="cramer")` we get **0.155**. For a 3x3 table, df = 2, so it's considered a **weak association** between the 2 categories. There's relationship, but not strong.

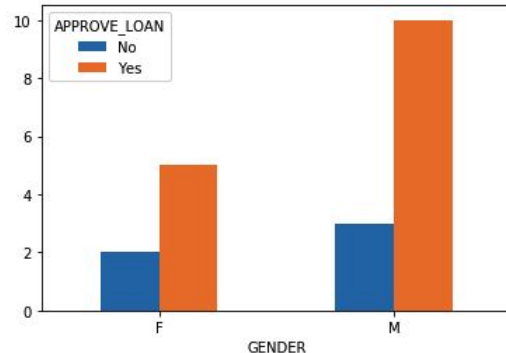
Two Categorical or Discrete Variables

Visualization – Grouped Chart

Multiple bars for each category are grouped together to facilitate comparison between different groups within each category (*one categorical in X, the count in Y, and the other categorical in color*).

Example: Company wants to automate the loan eligibility process based on customer details. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others. Let's look at the loan approval according to gender (M, F in our data).

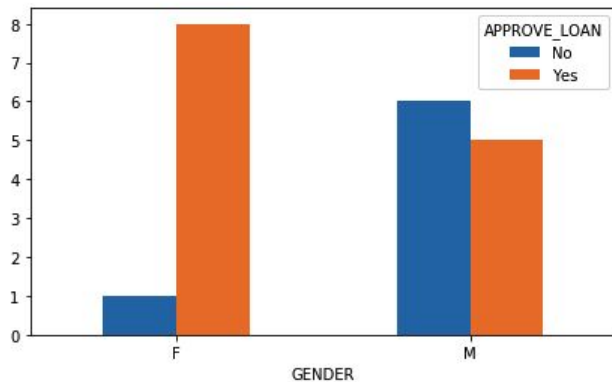
If the proportions of approvals (Yes/No) are similar between the genders, it suggests that gender might not be a significant factor influencing the loan approval decision.



Two Categorical or Discrete Variables

Visualization – Grouped Chart

Example: let's look at another loan approval according to gender (M, F in our data).



Here, the disparity in approval ratios is evident.

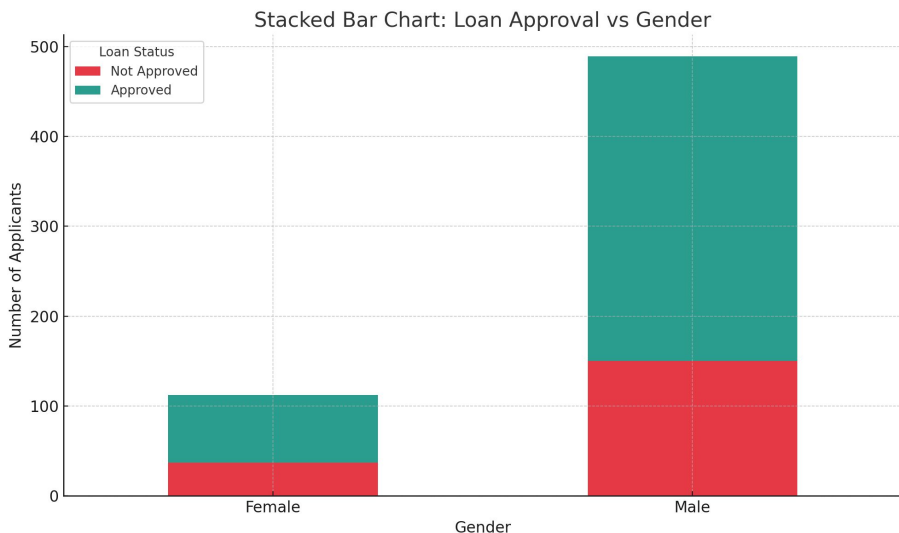
- *If you're female, your loan is more likely to be approved.*
- *On the other hand, for males, the odds of approval are roughly even.*

The data suggests a correlation between gender and loan approval rates.

Two Categorical or Discrete Variables

Visualization – Stacked Chart

Each bar is divided into multiple sub-categories, allowing for comparison of the overall category size as well as the proportions of its sub-categories.



Visually, the proportion of approvals seems relatively consistent between the two genders, with perhaps a slightly higher proportion for males.

This suggests that, while there might be some variation, gender might not be a significant factor in the loan approval process.

However, a deeper statistical analysis would be needed to confirm this observation.

Bivariate Analysis

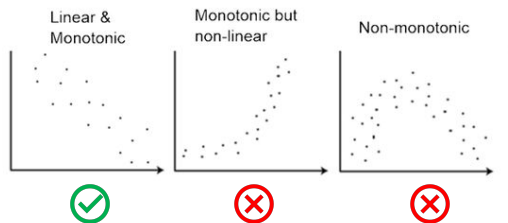
Numerical Continuous Variables

Two Numerical Continuous Variables

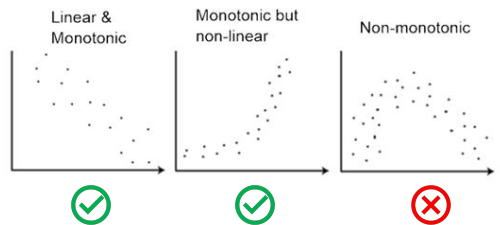
Correlation Coefficients

One of the primary tools used to measure the **strength** and **direction** of the **relationship** between two continuous numerical variables is the correlation coefficient.

Pearson's: for **linear relationships**.



Spearman's: for **monotonic relationships**.

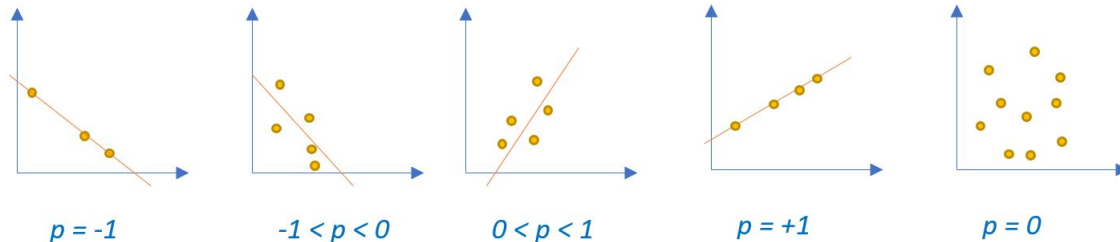


Two Numerical Continuous Variables

Correlation Coefficients

Pearson's: for **linear relationships**.

- **Positive values:** As one variable increases, the other also tends to increase.
- **Negative values:** As one variable increases, the other tends to decrease.
- The magnitude (absolute value) indicates the **strength**. Closer to 1 or -1 means a stronger relationship, while closer to 0 indicates a weaker relationship.



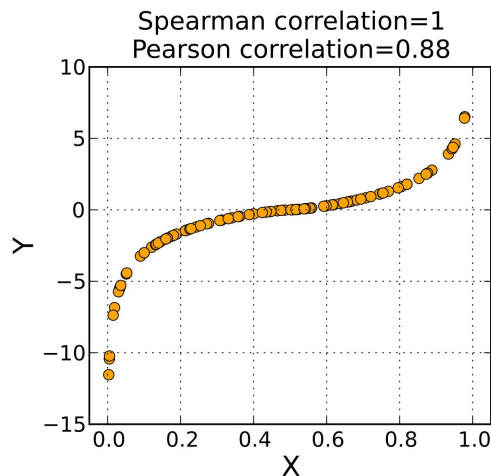
* Careful with outliers!

Two Numerical Continuous Variables

Correlation Coefficients

Spearman's: for **monotonic relationships**.

- **Positive values:** As one variable increases, the other also tends to increase, but not necessarily at a constant rate.
- **Negative values:** As one variable increases, the other tends to decrease, but not necessarily at a constant rate.



* Careful with outliers!

Two Numerical Continuous Variables

Correlation Coefficients

Davis, 1971 ^a		.70 or higher	Very strong association
		.50 to .69	Substantial association
		.30 to .49	Moderate association
		.10 to .29	Low association
		.01 to .09	Negligible association
Hinkle, Wiersma, & Jurs, 1979 ^{ab}	Correlation coefficients	.90 to 1.00	Very high correlation
		.70 to .90	High correlation
		.50 to .70	Moderate correlation
		.30 to .50	Low correlation
		.00 to .30	Little if any correlation
Hopkins (1997) ^a		.90 to 1.00	Nearly, practically, or almost: perfect, distinct, infinite
		.70 to .90	Very large, very high, huge
		.50 to .70	Large, high, major
		.30 to .50	Moderate, medium
		.10 to .30	Small, low, minor
		.00 to .10	Trivial, very small, insubstantial, tiny, practically zero

Two Numerical Continuous Variables

Correlation Coefficients and p-values

The correlation coefficient itself doesn't tell us whether the **observed relationship in the sample data is statistically significant in the population**.

We need to also check the p-value. *In this case, It tests the null hypothesis that there's no relationship between the two variables in the population – we will soon explain in detail about how this works.*

- **P-value < 0.05:** This suggests that the observed **correlation** in the sample is **statistically significant** and is unlikely to have occurred by random chance (*we reject the null hypothesis.*)
- **P-value > 0.05:** This suggests that the observed **correlation** in the sample is **not statistically significant** and could have occurred by random chance (*we fail to reject the null hypothesis*)
- **Note:** A p-value of 0.05 means that there is only 5% chance that results from your sample occurred due to chance. A p-value of 0.01 means that there is only 1% chance.

Two Numerical Continuous Variables

Correlation Coefficients – Example

In Eldoria, a local company is studying its famous coffee beans. They're tracking "Color Intensity" to gauge a bean's strength and using "Time of Day" to see when people drink their coffee. The "Energy Boost" measure shows the general power of a bean based on its color, while "Energy Boost over Time" focuses on the specific effects of the special "Golden Eldoria Bean" throughout the day. With this data, the company aims to better understand its product and serve its customers.

Is there a relationship between the color intensity of the beans and the energy boost they provide?

	Color Intensity	Energy Boost	Time of Day	Energy Boost over Time
0	9.45	8.80	22.55	0.66
1	3.55	2.34	6.79	14.84
...

Pearson Correlation Coefficient: 0.8317

Spearman Correlation Coefficient: 0.8299

Both correlation coefficients indicate a **strong positive association**. The values being close to each other and above 0.8 imply that **as the color intensity of the beans increases, the energy boost they offer also generally increases**.

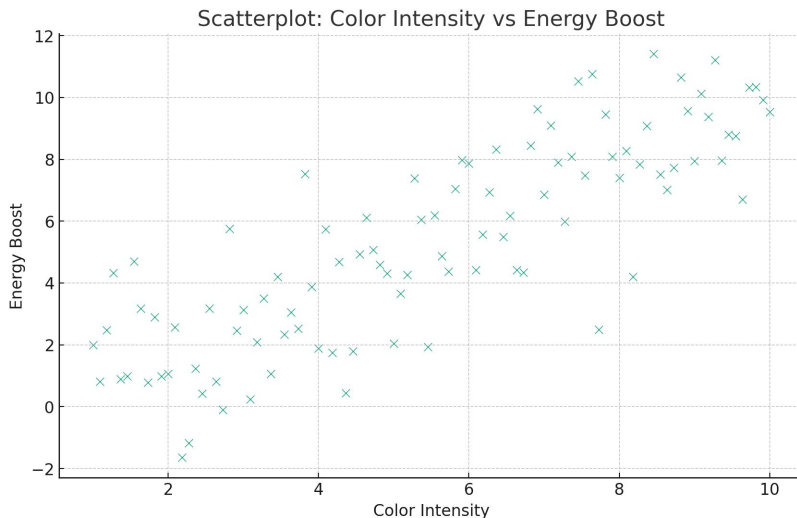
Two Numerical Continuous Variables

Visualization – Scatter Plots

Each point represents a single observation. The position of a point is determined by its values for the two variables being plotted (like coordinates).

Is there a relationship between the color intensity of the beans and the energy boost they provide?

As the color intensity increases, the energy boost generally increases as well, but with some variability.



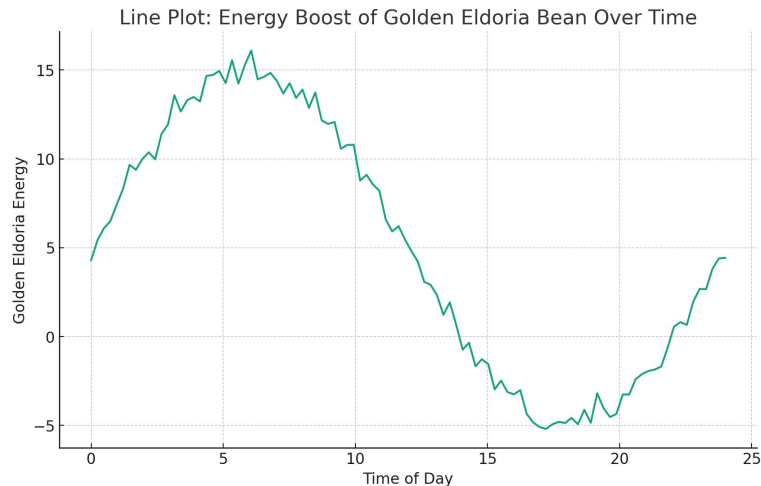
Two Numerical Continuous Variables

Visualization – Line Plots

A line plot displays information as a series of data points connected by straight line segments. It is often used to visualize a **trend in data over intervals of time** – hence its frequent use with **time series data**.

How does the energy boost from the "Golden Eldoria Bean" vary over the course of a day?

There's a noticeable pattern, likely sinusoidal, indicating peak energy boosts at certain times.



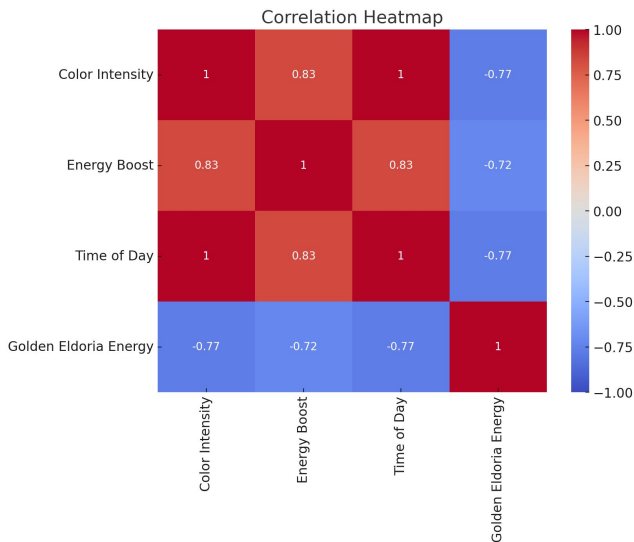
Two Numerical Continuous Variables

Visualization – Correlation Heatmaps

A correlation heatmap uses a color scale to represent the **correlation coefficients between pairs of variables in a matrix format**. It allows for a **quick visual assessment of relationships** (or lack thereof) between multiple variables.

How do different variables correlate with each other?

For instance, 'Color Intensity' and 'Energy Boost' have a high positive correlation, which was evident from the scatterplot as well.



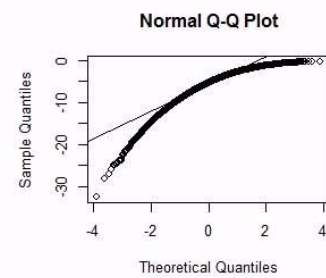
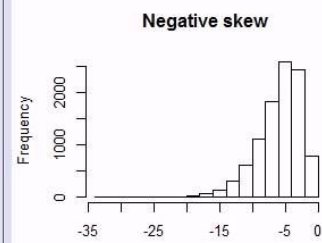
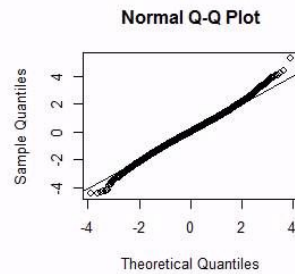
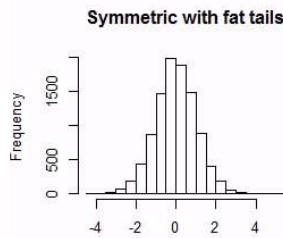
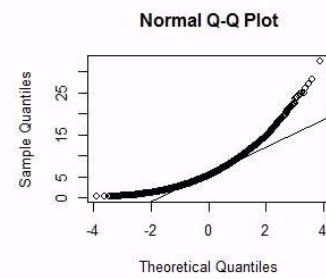
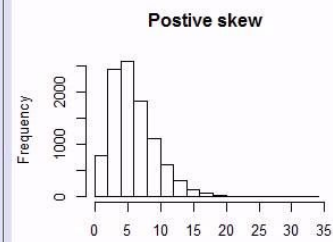
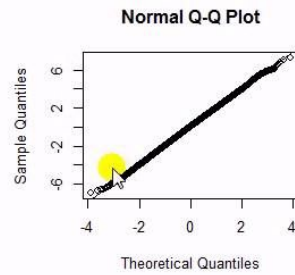
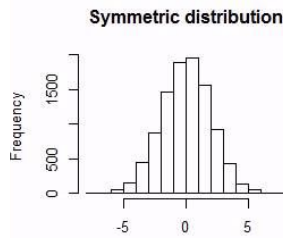
Two Numerical Continuous Variables

Visualization – QQ Plots

QQ plots, or Quantile-Quantile plots, are used to help assess if a dataset follows a certain theoretical distribution. Is done by plotting their quantiles against each other. If the two distributions being compared are similar, the points in the QQ plot will approximately lie on the line $y=x$.

- **Normality Testing:** One of the most common uses. In this case, the data quantiles are plotted against the quantiles of a standard normal distribution.
- **Comparing Distributions:** Besides the normal distribution, you can use QQ plots to see if your data fits other theoretical distributions like exponential, t-distribution, etc.

QQ - Plot

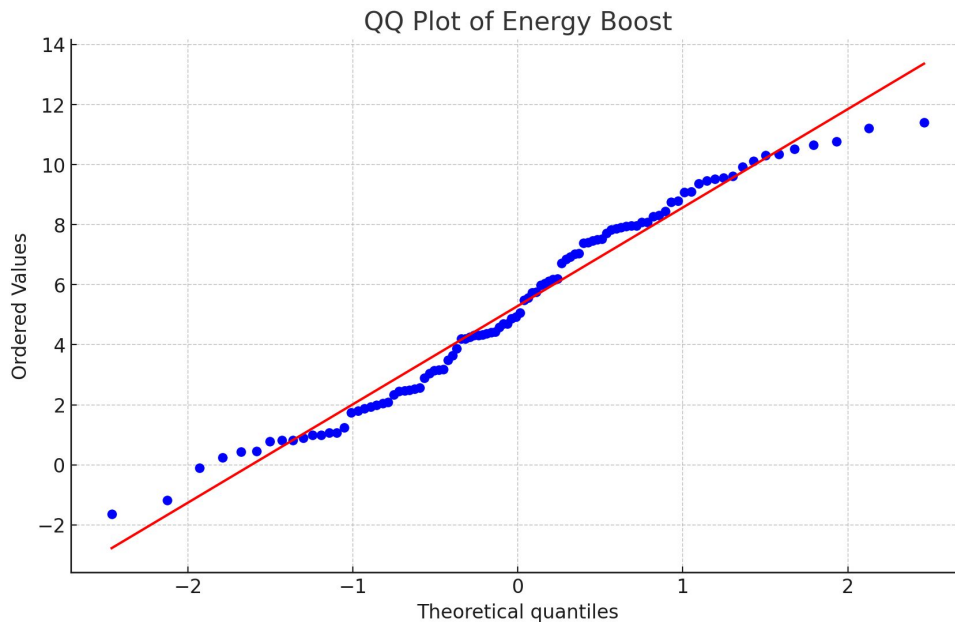


Two Numerical Continuous Variables

Visualization – QQ Plots

Is 'Energy Boost' data normally distributed?

If the data were perfectly normal, the points would lie on the straight line. The slight deviations, especially at the tails, suggest some differences from a perfect normal distribution.



Bivariate Analysis

Numerical Continuous VS

Categorical/Discrete Variables

Numerical Continuous VS Categorical Variables

Summary Statistics by Category

When you want to compare a numerical continuous variable with a discrete or categorical variable numerically, you're essentially trying to understand **how the continuous variable's values differ across the levels or categories** of the discrete variable.

- Calculate measures of centrality (like mean or median) and spread (like standard deviation or IQR) for the continuous variable within each category.
- For instance, if you're comparing exam scores (continuous) by grade level (categorical), you might compute the mean score for each grade.

You could also transform the continuous variable to categorical or vice-versa if it's what your analysis needs.

Numerical Continuous VS Categorical Variables

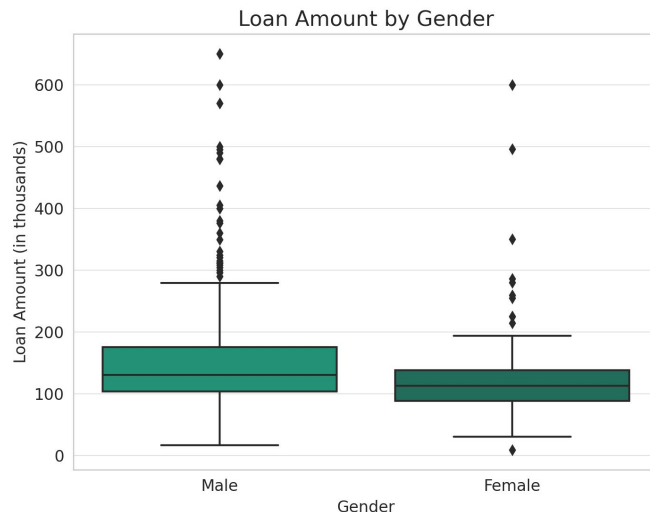
Visualization – Box plot

Show distributions of the numerical variable (axis Y) across categories (axis X).

Example: let's return to the loan approval case according to gender.

Is there any significant difference in the loan amounts between male and female requestors?

The median loan amount requested by both genders is relatively close, with males having a slightly higher median. Both genders have a few outliers, indicating some loans with unusually high amounts.



Numerical Continuous VS Categorical Variables

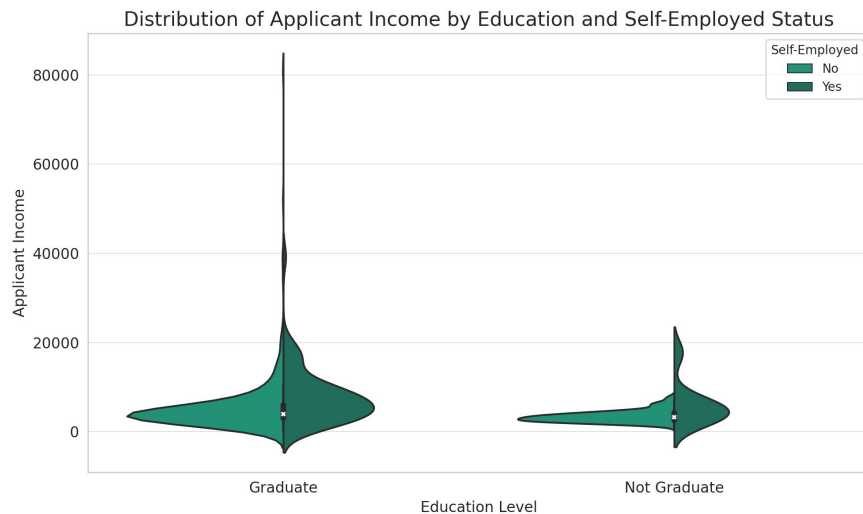
Visualization – Violin plots

Combine aspects of box plots and density plots.

How does the distribution of applicant income vary based on education and whether they are self-employed?

The broader sections of the violin plot represent higher density (more applicants fall within that income range). Graduates tend to have a wider income distribution than non-graduates.

For both graduates and non-graduates, self-employed individuals have a wider income distribution compared to those who are not self-employed. This might indicate that self-employed individuals can have highly variable incomes.



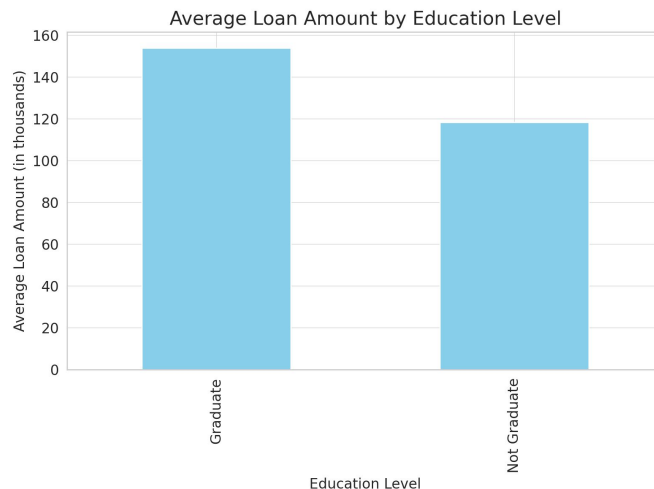
Numerical Continuous VS Categorical Variables

Visualization – Bar Charts

If in the **Y-axis instead of the count** or frequency for the categories, we show the **mean –or another measure of central tendency–** of the continuous variable for each **category (X-axis)**.

What is the average loan amount requested by applicants based on their education level?

Graduates, on average, request a slightly higher loan amount compared to non-graduates.



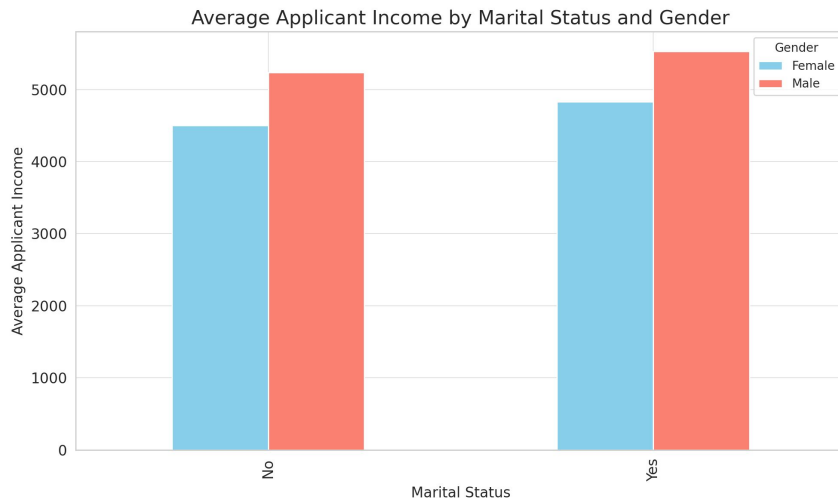
Numerical Continuous VS Categorical Variables

Visualization – Grouped Bar Charts

We can use grouped bar charts to visualize the relationship between a **categorical (X)** and a **continuous variable (Y)** –instead of count in Y–, differentiated by color by **another categorical** variable.

How does the average applicant income vary based on their marital status, differentiated by gender?

Both males and females who are married tend to have a higher average income compared to those who are not married. Males, on average, have a higher income than females, irrespective of their status.



Numerical Continuous VS Categorical Variables

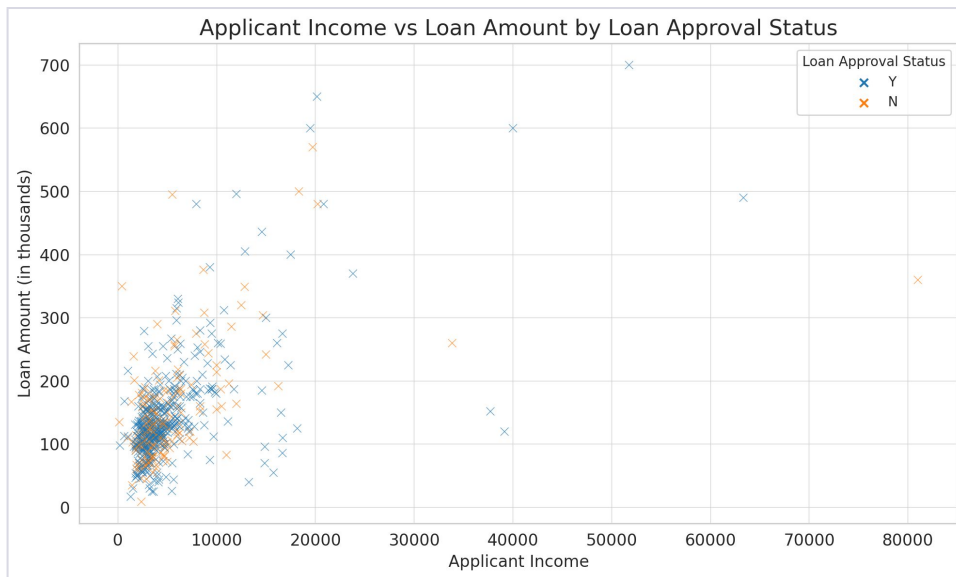
Visualization – Scatter Plots

We can use **a scatter plot** to visualize the relation between a continuous (X) with a continuous Y, **differentiated by color by a categorical variable**.

How does the applicant income relate to the loan amount requested, differentiated by loan approval status?

There's a general positive correlation, meaning as the applicant income increases, the loan amount requested tends to increase as well.

Both approved and denied loans are scattered throughout, but we do see a concentration of approved loans in the lower income and loan amount regions.



Numerical Continuous VS Categorical Variables

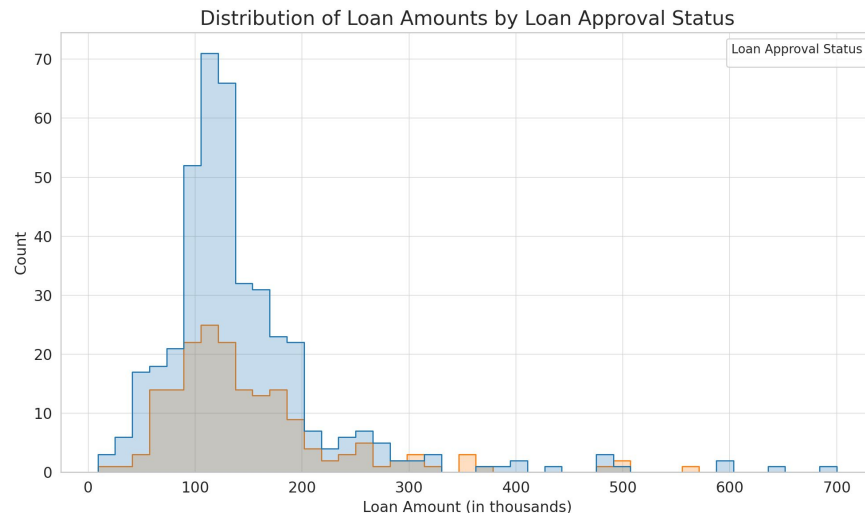
Visualization – Grouped Histogram

We can use **grouped histogram** to visualize the relation between a continuous (X) its count in Y, **differentiated by color by a categorical variable**.

How is the distribution of loan amounts requested by applicants, differentiated by their loan approval status?

Most loan applications, whether approved or denied, are concentrated in the lower loan amount range.

Approved loans are more frequent in the lower loan amount ranges, while denied loans have a more even distribution across various loan amounts.



Outliers

The background features a smooth gradient from light blue on the left to a mix of purple and pink on the right. A large, semi-transparent blue shape overlaps the right side of the image. Several small, colored dots (blue, orange, and white) are scattered across the background, adding a decorative, starry effect.

Outliers

What is an Outlier?

An outlier is an observation that appears to **deviate markedly from other members of the sample** in which it occurs. In simpler terms, it's a value considerably different from most of the other values in your dataset.



Causes of Outliers

- **Measurement or Input Error:** Mistakes can happen during data collection, recording, or entry.
- **Data Processing Error:** Incorrect calculations or data manipulation.
- **Natural Variations:** Genuine variations in data.
- **Intentional Outliers:** Purposely inserted anomalies to flag special attention (such as missing values).
- **Changes Over Time:** Evolving data values make old ones appear unusual.
- **Fraudulent or Malicious Activity:** Deceptive actions causing data irregularities.

Outliers

Techniques to Identify Outliers

1. Visual Inspection

- **Box Plots:** The whiskers in a box plot often extend 1.5 times the interquartile range (IQR). Data points outside this range can be considered outliers.
- **Histograms:** Outliers may manifest as isolated bars or peaks distant from the main data distribution.
- **Scatter Plots:** Outliers will appear as points that are far away from the cluster of other data points.

2. Statistical Methods

- **IQR Method:** Calculate the IQR (difference between 75th percentile and 25th percentile). Any data point outside $[Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR]$ can be considered as outliers.
- **Standard Deviation Method:** If a data point is more than n standard deviations away from the mean, it can be considered an outlier. Common values for n are 2 or 3.

3. Machine Learning

- Algorithms like DBSCAN, Isolation Forest, and One-Class SVM can detect outliers.

Outliers

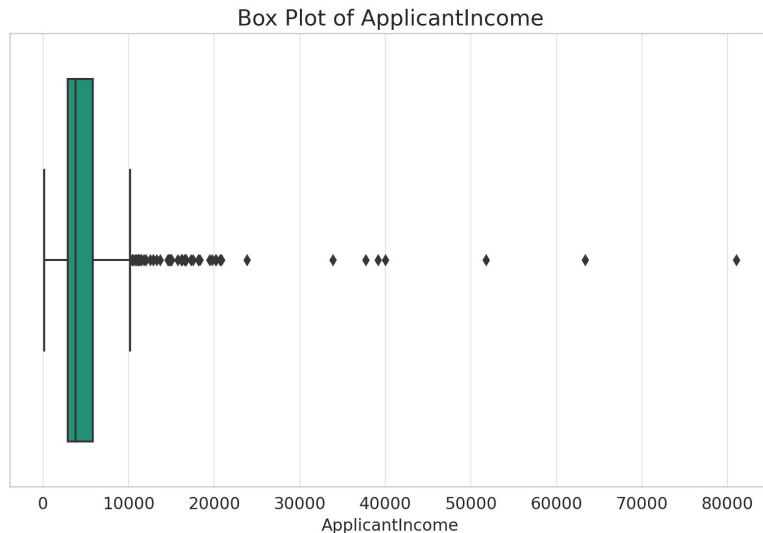
Techniques to Identify Outliers – Loan Approval Example

To identify outliers, we will focus on a numerical column. Let's choose *ApplicantIncome* for this purpose.

1. Visual Inspection

- **Box Plots**
- **Histograms**
- **Scatter Plots**

Outliers can often be identified as points outside the "whiskers" or ends of the box plot.



Outliers

Techniques to Identify Outliers – Loan Approval Example

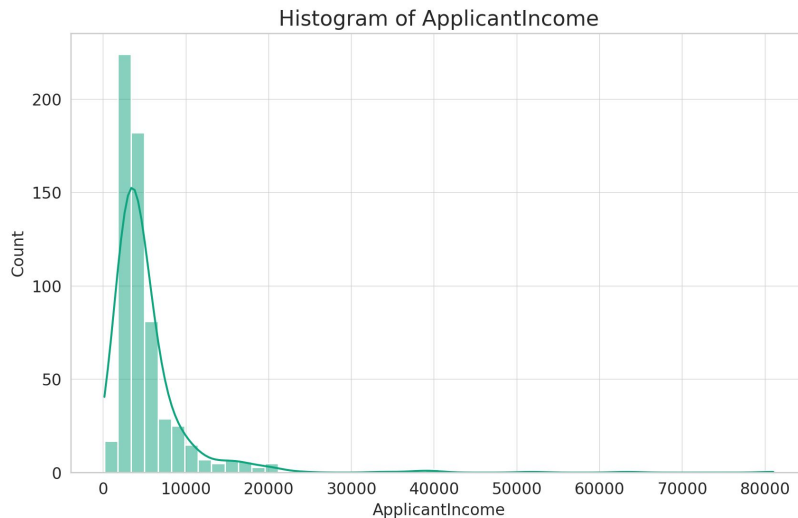
To identify outliers, we will focus on a numerical column. Let's continue with *ApplicantIncome*.

1. Visual Inspection

- Box Plots
- Histograms
- Scatter Plots

A histogram represents the distribution of a dataset. Outliers can sometimes be visualized as isolated bars far from the central data cluster.

The histogram for ApplicantIncome shows a right-skewed distribution. Most of the applicants have an income in the lower range, but there are a few applicants with a much higher income, which can be seen as the long tail on the right side. These can be potential outliers.



Outliers

Techniques to Identify Outliers – Loan Approval Example

For a Scatter Plot, we need two numeric continuous variables. Let's create a scatter plot comparing *ApplicantIncome* against *LoanAmount* to identify potential outliers.

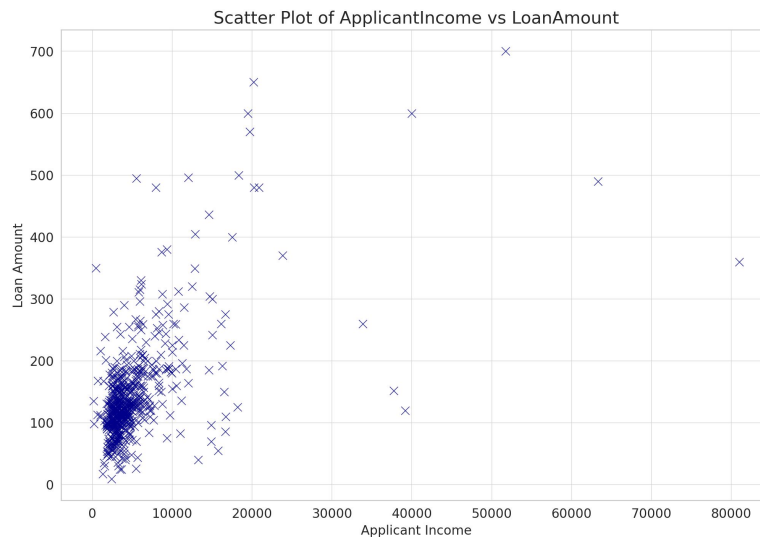
1. Visual Inspection

- Box Plots
- Histograms
- Scatter Plots

By using scatter plots, you can visually inspect and identify potential outliers as **isolated data points** when **comparing two numerical variables**.

The scatter plot illustrates the relationship between ApplicantIncome and LoanAmount. Most data points are clustered around the lower income and loan amount ranges, which is expected.

However, you can also notice a few points that are positioned away from the main cluster, especially on the right side of the plot. These points represent applicants with higher incomes and potentially larger loan amounts. Such points can be considered outliers in the context of this relationship, as they deviate from the general trend observed in the data.

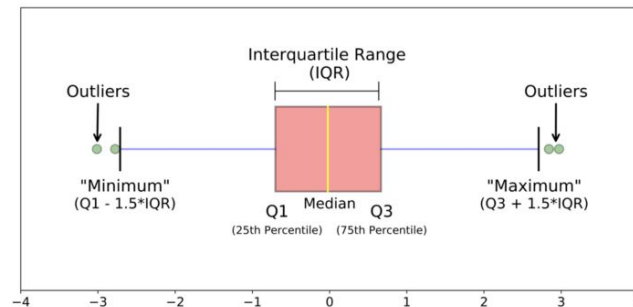


Outliers

Techniques to Identify Outliers – More Detail

2. Statistical Methods

- **IQR Method (Tukey's Method)**
- **Standard Deviation Method**



Same as through box plots but with numerical calculations: we calculate the IQR ($Q3 - Q1$). Any **data point outside the range** $[Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR]$ can be considered as **outliers**.

The 1.5 multiplier (**scale**) is standard for outlier detection. Alternatives include:

- **1.0:** Aggressive, will identify **more points as outliers**. It might be used when we want to be more conservative and ensure we're capturing potential anomalies, especially in datasets with fat tails.
- **2.0 or 3.0:** More forgiving thresholds, **reduces** the number of **identified outliers**. This might be used in cases where the data is known to have wide variations but these are still considered typical.

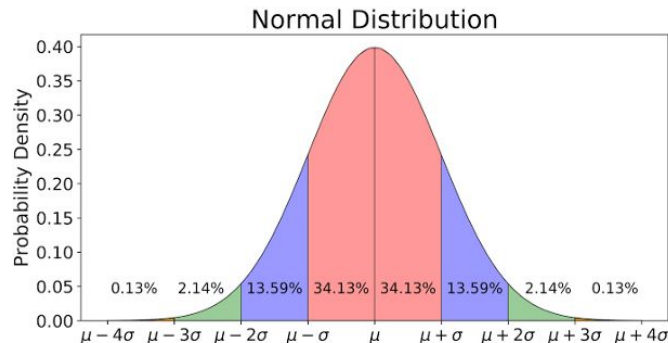
This number, the *scale*, depends on the distribution followed by the data.

Outliers

Techniques to Identify Outliers – More Detail

2. Statistical Methods

- IQR Method (Tukey's Method)
- Standard Deviation Method



This method is used if our **data is normally distributed**. If it's not our case, we then use the IQR method, or transform our data (log, square root, or Box-Cox – will talk about this later) to make it more normal-like before applying the standard deviation method.

If a data point is more than **n standard deviations away** from the mean, it can be considered an outlier. A common threshold is **n=3** since for a dataset that is approximately **normally distributed**:

- About **68%** of the data lies within **1 standard deviation of the mean**.
- About **95%** lies within **2 standard deviations**.
- About **99.7%** lies within **3 standard deviations**.

Outliers

Techniques to Identify Outliers – More Detail

Why 1.5?

The 1.5 multiplier in the IQR outlier detection is inspired by the normal distribution characteristics, although the IQR method doesn't require the data to be normal.

$$\begin{aligned}\text{Upper Bound} &:= Q3 + 1.5 * IQR \\ &= Q3 + 1.5 * (Q3 - Q1) \\ &= 0.675\sigma + 1.5 * (0.675 - [-0.675])\sigma \\ &= 0.675\sigma + 1.5 * 1.35\sigma \\ &= \mathbf{2.7\sigma}\end{aligned}$$

Same for lower bound

Using the 1.5 multiplier, the IQR method flags data points beyond **2.7 σ** from the mean as outliers. This threshold is close to the normal distribution's **3 σ** rule for outliers, aligning the IQR method with common outlier detection practices in normally distributed data.

Note: Q1 (the first quartile) represents the 25th percentile. This means 25% of the data is below this point. When you look up this percentile in a standard normal distribution table or use statistical software to get the z-score, it corresponds to approximately -0.675. Same reasoning for Q3 and +0.675.

Outliers

What to do with Outliers?

- **Investigate:** Before making a decision, understand why the outlier exists. It might offer a new insight or indicate an error.
- **Keep:** If the outlier is a result of natural variation and is important for analysis, keep it.
- **Remove:** If it's due to an error or it biases the analysis, consider removing it.
- **Transform:** Apply transformations like log or square root to dampen the impact of the outlier.
- **Separate Analysis:** Analyze outliers separately if they represent a different group or phenomenon.
- **Using Different Analysis Techniques:** You could also use different statistical techniques or Machine Learning models that are not as much impacted by the presence of outliers.

Why visualizations

Visualization Techniques

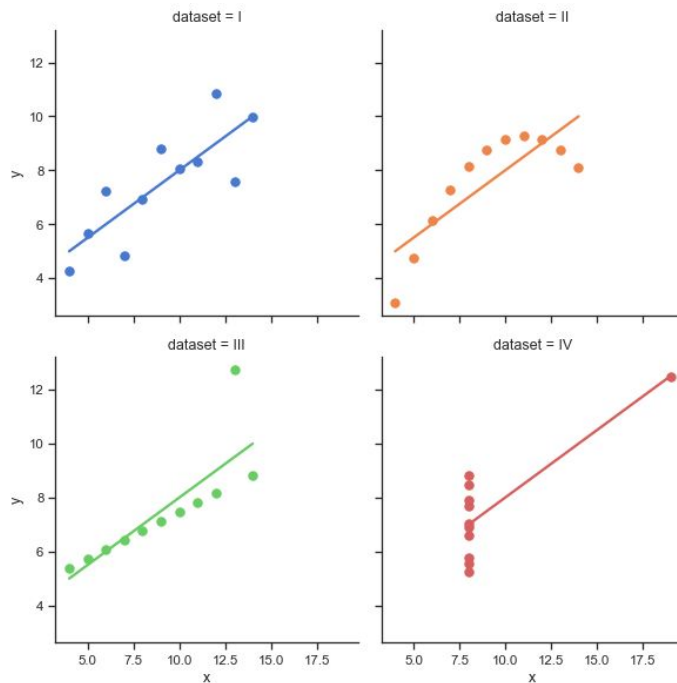
Example: Anscombe's quartet

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

It comprises four different datasets that have nearly identical measures of centrality (mean) and dispersion (variance) for both x and y variables.

Visualization Techniques

Example: Anscombe's quartet



Summary

Recap – Summary

- **Univariate Analysis:** Focuses on a **single variable**
 - **Categorical** variables:
 - **Frequency** tables. Counts and proportions.
 - Visualizations: **Bar charts, pie charts**
 - **Numerical** variables:
 - Measures of centrality:
 - **Mean, median, mode**
 - Measures of dispersion:
 - **Variance, standard deviation, minimum, maximum, range, quantiles**
 - Shape of the Distribution:
 - **Symmetry and kurtosis**
 - Visualizations: **Histograms, box plots**

Summary

- **Bivariate Analysis:** Examines the **relationship** or **association** between **two variables**.

Type of Variables (VS)	Categorical (incl. Discrete numerical)	Continuous
Categorical (incl. Discrete numerical)	<ul style="list-style-type: none">• Crosstab• Chi-square tests• Cramér's V• Stacked• Grouped bar charts• Frequency heat maps	<ul style="list-style-type: none">• Violin Plots• Bar Charts (Shows the mean -or another measure of central tendency- of the continuous variable for each category)• Side by side Box Plots
Continuous	<ul style="list-style-type: none">• Violin Plots• Bar Charts (Shows the mean -or another measure of central tendency- of the continuous variable for each category)• Side by side Box Plots	<ul style="list-style-type: none">• Correlation coefficients, Covariance• Regression analysis, Coefficient of Determination• Scatter plots• Line plots• Correlation Heatmaps• QQ Plot

Note: numerical techniques are in this colour.

Summary

Type of Variables (VS)	Method/Technique	One-Line Description
Categorical (incl. Discrete numerical) VS Categorical	Crosstab	A table showing the frequency of occurrences for combinations of two categorical variables.
	Chi-square tests	Tests the independence of two categorical variables by comparing observed frequencies to expected frequencies.
	Cramér's V	Measures the strength of association between two categorical variables.
	Stacked or grouped bar charts	Visualizes the frequency or proportion of categories between two categorical variables.
	Frequency heat maps	Displays frequencies using color gradients for combinations of two categorical variables.
Categorical VS Continuous	Violin Plots	Combines a box plot with a kernel density plot to show the distribution of a continuous variable for each category .
	Bar Charts	Shows the mean (or another measure of central tendency) of the continuous variable for each category .
	Side by side Box Plots	Displays the distribution of the continuous variable for each category, showing quartiles and potential outliers .
Continuous VS Continuous	Correlation coefficients, Covariance	Quantifies the strength and direction of the relationship between two continuous variables.
	Regression analysis, Coefficient of Determination	Describes the relationship between two continuous variables and measures how well the regression line fits the data.
	Scatter plots	Plots individual data points based on their values for two continuous variables to visualize relationships .
	Line plots	Connects individual data points with lines , typically used to represent sequences or time series data.
	Correlation Heatmaps	Visualizes correlation coefficients between pairs of continuous variables using color gradients.
	QQ Plot	Compares the quantiles of a variable's distribution to the quantiles of a standard normal distribution .



Bonus

Normal Distribution

Checking if data is normally distributed

Visual Inspection:

- **Histogram:** A bell-shaped curve in a histogram is indicative of a normal distribution.
- **Q-Q Plot:** In this plot, the quantiles of your data are plotted against the quantiles of a normal distribution. If the data is normally distributed, the points should roughly lie on the $y=x$ line.
- **Box Plots:** The symmetry of a box plot can give hints about data normality.

Statistical Tests:

- **Shapiro-Wilk Test:** This is a popular test for normality. A low p-value (typically $p < 0.05$) indicates that the data is not normally distributed.
- **Kolmogorov-Smirnov Test:** This test compares the cumulative distribution of your data to a normal distribution. Again, a low p-value suggests non-normality.

Descriptive Statistics:

- **Skewness and Kurtosis:** Skewness measures the asymmetry of the data distribution, while kurtosis measures the "tailedness". For a normal distribution, **skewness** should be **close to 0** (indicating symmetry), and **kurtosis** should be **close to 3**. Should be used in conjunction with other methods.

Normal Distribution

Transforming Data to Be Normally Distributed

Transforming data to be approximately normal can aid in statistical analysis and modeling.

Log Transformation:

- Useful for data that shows exponential growth, like population or financial data.
- Use when data is right-skewed.

Square Root Transformation:

- Moderates the impact of extreme values.
- Suitable for data with mild skewness.

Box-Cox Transformation:

- Requires positive data values.
- Automatically determines the best power transformation.

Normal Distribution

Transforming Data to Be Normally Distributed

After Transformation:

- **Re-assess Distribution:** After applying a transformation, visually assess the distribution again using histograms and Q-Q plots.
- **Statistical Testing:** Shapiro-Wilk or Kolmogorov-Smirnov tests can be used to statistically assess normality.
- Remember to **reverse** transformations (when needed) for interpretation.

Always consider the underlying **reasons** for any non-normality, as transformations might not always be the best solution.

Normal Distribution

Central Limit Theorem

The CLT states that, regardless of the shape of the underlying population, the **sampling distribution of the mean** will **approximate** a **normal distribution** as the **sample size grows larger** ($n > 30$), assuming all samples are identical in size and are randomly sampled.

1. **Large Sample Size & Individual Data Points:** Even with a large sample, the distribution of individual data points could still be non-normal. For instance, a dataset with millions of data points could still be heavily skewed or have extreme kurtosis.
2. **Large Sample Size & Averages of Samples:** If you're taking multiple samples from a population and calculating their averages, the distribution of those averages tends to be normal due to the CLT, even if the underlying population is not normal.
3. **Practical Implications:** While the CLT is powerful, remember that many statistical tests and methods assume that the individual data points (not their means) are normally distributed. So, you can't bypass these assumptions simply because you have a large dataset.