How to use this guide

1. Download the Titanic training file (train.csv) from Kaggle (use the **train.csv** in the Kaggle "Titanic — Machine Learning from Disaster" dataset).
2. Open the CSV in Excel. Work on a **copy** (File → Save As → `EDA_Titanic_YourName.xlsx`).
3. For reproducibility convert the data into a Table: select any cell → **Insert** → **Table** → ensure "My table has headers" is checked → **OK**. This makes ranges stable for PivotTables and charts. (All video steps assume a Table or contiguous range.)
4. If you don't have it, enable the Data Analysis ToolPak: **File → Options → Add-ins → Manage: Excel Add-ins → Go → check Analysis ToolPak → OK**

Part 1 — Basic Numerics (Descriptive Statistics)

Video: *Exploratory Data Analysis With Excel - Part 1 - Basic Numerics*.

Objective

Get numeric summary statistics (count, mean, median, std dev, min, max, skewness) for numeric columns such as **Age** and **Fare**.

Exact step-by-step (clicks & options)

1. Select the sheet with the Titanic table. Ensure Age and Fare look numeric. If they display as text, convert them (select column → Data → Text to Columns → Finish) or use `VALUE()` in a helper column and paste values. (Video checks types visually.)
2. Enable Data Analysis ToolPak if not present: **File → Options → Add-ins → Manage: Excel Add-ins → Go → check "Analysis ToolPak" → OK**.
3. **Data → Data Analysis → Descriptive Statistics → OK**.
    - **Input Range:** click and select the numeric column (include header if using "Labels in first row").
    - **Labels in first row:** check if you included headers.
    - **Output Range:** pick a blank cell (or choose New Worksheet Ply).
    - **Summary statistics: check** this box (required to get Mean, Std Dev, etc.).
4. Click **OK** → Excel generates a table with Count, Mean, Std Dev, Min, Max, Skewness, Kurtosis, etc.
5. Repeat for **Age** and **Fare** (or select both columns at once as Input Range if you want simultaneous summary).

What you should record

- Count (how many non-blank values)
- Mean & Median (central tendency)
- Std Dev & Variance (spread)
- Min & Max (range)
- Skewness (shape of distribution) — helpful to decide visualization type

Concept explained

- **Mean vs Median:** mean is average, median is middle value; median is robust to outliers.
- **Std Dev:** average distance from mean; large std dev = widely spread values.
- **Skewness:** positive skew means long tail to right (e.g., Fare often has large right tail because a few passengers paid very high fares).

Quick checks / common pitfalls

- Blanks in Age reduce count and will affect mean — note how many blanks the Descriptive Statistics output shows. The video calls out missing Age values for further handling.

Part 2 — Basic Categoricals (Counts, Cross-tabs)

Video: *Exploratory Data Analysis With Excel - Part 2 - Basic Categoricals*.

Objective

Get frequency counts and contingency tables for categorical variables: **Survived**, **Sex**, **Pclass**, **Embarked** (e.g., how many males vs females, survival counts by sex, etc.).

Step-by-step (counts with formulas + pivot)

A. Quick formula counts (to replicate exactly what video shows):

- Count males: `=COUNTIF(Table1[Sex],"male")`
- Count survivors: `=COUNTIF(Table1[Survived],1)` — in Titanic dataset Survived = 1 indicates survived.
- Count female survivors: `=COUNTIFS(Table1[Sex],"female",Table1[Survived],1)`

B. PivotTable (video uses pivots to create tidy cross-tabs):

1. Click any cell in the Table. **Insert → PivotTable**. In dialog choose **New Worksheet → OK**.
2. In **PivotTable Fields** pane:
    - Drag **Sex → Rows**.
    - Drag **Survived → Columns**.
    - Drag **PassengerId** (or any unique ID) → **Values** (it shows **Count of PassengerId**). This produces counts of passengers by Sex × Survived.
3. To show survival proportions by sex: click **Count of PassengerId** in Values → **Value Field Settings → Show Values As → % of Row Total → OK**. Now each row shows the proportion of survivors/non-survivors within the sex.
4. Repeat pivot with **Pclass** in Rows and **Survived** in Columns to get survival by class.

Concept explained

- **Categorical variables** are discrete labels — we use counts and proportions to summarize them.
- **Contingency tables** (pivot tables) show joint distributions: e.g., how survival depends on sex and class.

Logical thinking / insights students should look for

- Compare survival % between females and males — the dataset historically shows females had much higher survival rates. (See dataset analysis citations below.)

Part 3 — Histograms

Video: *Exploratory Data Analysis With Excel - Part 3 - Histograms*.

Objective

Visualize distribution of numeric columns (**Age**, **Fare**) and detect skew, central mass, and outliers.

Step-by-step (ToolPak histogram - video method)

Method shown in the video (Data Analysis → Histogram):

1. Create bin cutoffs in a helper column (for Age example): type `0, 10, 20, 30, 40, 50, 60, 70, 80`. These are the upper limits of bins.
2. **Data → Data Analysis → Histogram → OK**.
    - **Input Range:** select Age column (exclude header if not using labels).
    - **Bin Range:** select your helper bin column.
    - **Output Range:** choose a blank area or new worksheet.
    - **Chart Output: check** this box (makes Excel build the histogram chart).
3. Click **OK**. Excel outputs frequency counts and a histogram chart.

Alternative (newer Excel versions) — Insert Chart method (if video demonstrates it):

- Select Age column → **Insert → Insert Statistic Chart → Histogram** → format bin width with right-click → **Format Axis** → set **Bin width** or number of bins.

Concept explained

- **Histogram bins** group continuous values; wider bins smooth noise, narrow bins show detail.
- **Skew and outliers:** histograms show whether data is symmetric or skewed — e.g., Fare is typically right-skewed due to a few expensive tickets.

Thinking prompts / insights to record

- Count how many ages are missing — missing Age values affect the histogram. Consider whether to impute or exclude for subsequent analysis.

Part 4 — Box Plots (Box & Whisker)

Video: *Exploratory Data Analysis With Excel - Part 4 - Box Plots.*

Objective

Use box plots to show the five-number summary (Min, Q1, Median, Q3, Max) and detect outliers for Age and Fare, and compare groups (e.g., Age distribution for survivors vs non-survivors).

Step-by-step

1. If plotting a single variable: select the numeric column (Age) → **Insert** → **Insert Statistic Chart** → **Box and Whisker** → Excel draws a box plot.
2. To compare groups (Age by Survived): create two side-by-side columns — Age of Survivors and Age of Non-Survivors. Easiest method (as shown in the video):
    - Insert a PivotTable: put **Survived** in Columns, **Age** in Values but set to **Average** (or use **Field Settings** to get multiple summary statistics), then copy the Age columns out into regular columns (values only) arranged side-by-side.
    - Select these two columns → **Insert** → **Box & Whisker** (Excel will treat each column as a separate series and show them side by side).
3. Format: click chart → **Chart Elements** → turn on Data Labels if needed; right-click axis → **Format Axis** to set scale consistent across comparisons.

Concept explained

- **Box plot** visualizes distribution and outliers; the box covers Q1–Q3, the line inside is the median, whiskers extend to non-outlier extremes, and stand-alone points are outliers.
- Useful to compare central tendency and spread across groups (e.g., survivors vs non-survivors).

Insight to look for

- Does the median Age differ between survivors and non-survivors? Are there more extreme older passengers in one group? This helps form hypotheses about age and survival.

Part 5 — Bar Charts

Video: *Exploratory Data Analysis With Excel - Part 5 - Bar Charts*.

Objective

Show categorical counts or aggregated numeric summaries (e.g., counts of survivors by Pclass, or average fare by Pclass) using bar/column charts.

Step-by-step (from Pivot / summary table)

1. If you already have a PivotTable (e.g., Rows = Pclass, Columns = Survived, Values = Count of PassengerId) click inside the pivot and choose: **PivotTable Analyze → PivotChart** → pick **Clustered Column (or Bar) → OK**. The pivot chart links to the pivot.
2. If using a small summary table: select the two-column summary (e.g., Pclass vs Count) → **Insert → Column or Bar Chart → Clustered Column**.
3. Format: Add Chart Title (click title → type), axis titles (Chart Elements → Axis Titles), and Data Labels (Chart Elements → Data Labels). Use **Chart Tools → Format** to set number formatting on axis (e.g., integer counts).

Concepts explained

- Use bar charts for comparing discrete categories. Always label axes and include units. Bar length represents counts or aggregated value (sum/average).

Insights to record

- Which class has the highest survival count and rate? Compare counts (raw) to percentages — percentages show survival *rate* and are often more informative than raw counts.

Part 6 — Scatter Plots

Video: *Exploratory Data Analysis With Excel - Part 6 - Scatter Plots*.

Objective

Examine relationships between two numeric variables — e.g., **Age vs Fare** — to see correlation, clusters, and outliers.

Step-by-step

1. Prepare two columns (no blanks) — e.g., Age in column A and Fare in column B. If there are blanks, filter or remove them for plotting.
2. Select both columns → **Insert → Charts → Scatter (X, Y)** → choose **Scatter with only markers**.
3. Add a trendline (if shown in the video): click any data point → **Chart Elements (green +) → Trendline → Linear** or **Right-click data series → Add Trendline → select Linear**. If the video displays R-squared or equation: open **Trendline Options → Display Equation on chart** and **Display R-squared**.
4. Add axis titles via Chart Elements → **Axis Titles** → Type "Age" and "Fare".

Concept explained

- **Correlation** measures the strength/direction of linear relationship. Trendline slope and R-squared help summarize it. Scatter plots are great to spot outliers (very high fares) and clusters by region or class if you color code (video may keep it simple).

Insight to consider

- Do older passengers pay different fares? Is there a cluster of low-fare passengers? Outliers (very high fares) will appear far from the main cloud and can influence average Fare.

Part 7 — Treemap Charts

Video: *Exploratory Data Analysis With Excel - Treemap Charts*.

Objective

Show hierarchical categorical distribution — e.g., how counts break down by **Pclass → Sex** (size of rectangle = count).

Step-by-step

1. Build a summary table for the hierarchy. The video shows doing this via a PivotTable: put **Pclass** in Rows, **Sex** in Columns, and **PassengerId** in Values (Count). Then convert the pivot into a 3-column flat table: Pclass | Sex | Count (copy → Paste Values).
2. Select the 3-column summary → **Insert → Hierarchy Charts → Treemap**.
3. Turn on Data Labels (Chart Elements → Data Labels) and format so labels show counts or percentages (right-click labels → Format Data Labels → select Value or Percentage).
4. Resize and give a clear title (e.g., "Passengers by Class and Sex").

Concept explained

- **Treemap** visualizes hierarchical proportions as nested rectangles — area corresponds to magnitude. It's useful to see relative sizes at both levels (class and sex).

Thinking prompt

- Use treemap to spot which subgroups dominate (e.g., 3rd class males may be the largest rectangle). Ask: how do these absolute counts compare to survival rates? (Absolute size ≠ better outcome)

Concepts across the series — plain explanations (so students truly understand)

1. **Table vs Range:** Tables auto-expand and are easier for PivotTables and structured references.
2. **Descriptive statistics:** mean, median, mode, std dev, skewness, kurtosis — used to summarize the numeric variable behavior.
3. **Missing Data:** columns like *Age* and *Cabin* often have missing values — these must be noted and handled later (imputation or exclusion). The videos point to Age as requiring attention.
4. **Categorical summarization:** `COUNTIF`, `COUNTIFS`, and PivotTables convert raw labels into meaningful counts and rates.
5. **PivotTable mechanics:** Rows, Columns, Filters, Values — changing these changes the view; **Value Field Settings** controls summary function (Count, Sum, Average) and display (% of Row/Column/Grand Total).
6. **Grouping (Pivot):** Right-click a field like Age in the pivot → **Group** (set start/end and interval) to create age bins. Useful for pivot histograms without the ToolPak.
7. **Charts and interpretation:** choose chart type to match data: histograms/boxplots for distribution, bars for category comparison, scatter for relationships, treemap for hierarchical proportions.
8. **Outliers and skew:** high Fare outliers and missing Ages influence means and charts — use medians and boxplots for robust summaries.

Data insights & logical narrative (so students *think* alongside the analysis)

(These are the typical, reproducible outcomes when you run the EDA steps on the Kaggle Titanic train.csv. I cite EDA/Kaggle sources for the observed facts.)

1. **Overall survival rate:** ~38–62% depending on data slicing; the train.csv commonly shows ~38% survivors (varies by data split). Use `=AVERAGE(Table1[Survived])` to find the exact value in your file.
2. **Gender is highly predictive:** females have a substantially higher survival rate than males (commonly ~70% of females survived vs ~20–30% of males in typical EDA). In pivots you will see the female row dominated by survivors. This is visible in the categorical pivot in Video 2.
3. **Passenger class matters:** First-class passengers enjoyed much higher survival rates than 3rd class. Pivot and bar chart comparisons will show Pclass = 1 has the highest proportion of survivors.
4. **Age patterns:** Age distribution is concentrated in adults (20–40); children are present but fewer. Box plots for survivors vs non-survivors may show slightly different medians and spread — children sometimes survived at higher rates (policy "women and children first" historically influenced outcomes). Use grouped boxplots/histograms to visualize this.
5. **Fare is right-skewed:** Histogram and boxplot for Fare will show a long right tail (a small number of very high fares)—these are outliers that affect mean fare. Use median as robust central tendency.
6. **Missing values:** *Cabin* often has many missing values; *Age* has some missing entries — note counts from Descriptive Statistics. This is highlighted in the video series as important for later steps.

How students should document findings (workbook checklist)

On completion place the following sheets in your workbook (video follows a similar sheet organization):

- `Data` (original table) — leave untouched.
- `Descriptives` — outputs from Data Analysis ToolPak for Age and Fare.
- `Categorical_Summaries` — pivot(s) showing Sex × Survived and Pclass × Survived.
- `Histograms` — histograms for Age and Fare (with bin table).
- `Boxplots` — box & whisker charts for Age and Fare and grouped comparisons.
- `Charts` — bar charts, scatter, and treemap.
- `Insights` — short bullet list of 6–8 observations (what we listed above) with one-line evidence (e.g., "Females: 68% survival — pivot shows X of Y").

(These sheet names match the workflow shown across the video series.)

Student thinking prompts to build analytical reasoning (use while doing each step)

- After Descriptive Statistics: *Which measure (mean or median) better represents Age/Fare and why?*
- After Categorical pivots: *Is a higher raw count the same as a higher survival rate? Why or why not?*
- After Histogram/Boxplot: *Do outliers change your interpretation? Should you remove them or report them?*
- After Scatter: *Is the relationship linear or noisy? Would correlation alone prove causation?*
- After Treemap: *How does absolute count (treemap area) compare to survival proportions?*