Impact of working remotely on social wellbeing and productivity

Amaury Chartier Valentine Salvat Thomas Blondel
Elisa Rigazzio

2026-10-12

This study explores how remote work affects productivity, social well-being, and job satisfaction. Using the 2020 NSW Remote Working Survey, we applied descriptive statistics and correlation analysis to quantify attitudes toward remote work. Findings show mixed productivity outcomes: most employees report gains up to 50%, fewer others declines. Collaboration indicators remain neutral, while reduced commuting significantly increases personal and family time, supporting better work–life balance. Most respondents prefer more remote work, suggesting higher satisfaction overall. Limitations include self-reported data and cross-sectional design. Results indicate organizations should maintain flexible work arrangements and support social engagement strategies.

! Interpretation > Visualization

Plots alone won't earn you a good grade. What matters most is **interpreting your findings**.
For every result you present:

- **Explain** what the data shows
- **Interpret** what it means for your research questions
- **Discuss** implications and connect to domain knowledge

Quality of insight > Quantity of plots. Your instructors can read plots—show them you *understand* what the data reveals.

Quarto Guide (Remove After)

Two Report Writing Options

Option 1 - Modular (Recommended for Teams):

- Each section in a separate `.qmd` file in `report/sections/`
- Files prefixed with `_` (e.g., `_introduction.qmd`) are auto-included
- Better for collaboration (fewer merge conflicts)
- Render `report.qmd` to build the complete report

Option 2 - Single File:

- Write everything in `report.qmd`
- Simpler but harder to collaborate
- Delete `report/sections/` folder if using this approach

See Quarto includes documentation for details.

What Are Code Chunks?

Code chunks are blocks of executable code embedded in your Quarto document. They run when you render and include their output (plots, tables, results) in your report.
Basic Syntax:

- Start with three backticks followed by the language: ```` ```python ````
- Write your code
- End with three backticks: ```` ``` ````

Example:

```python
import pandas as pd
data = pd.DataFrame({'x': [1, 2, 3], 'y': [4, 5, 6]})
print(data)
```

Chunk Options (use `#| option: value` at the top):

- `#| echo: false` - Hide code, show only output
- `#| eval: false` - Show code but don't run it
- `#| output: false` - Run code but hide output
- `#| warning: false` - Suppress warning messages
- `#| fig-cap: "My Plot"` - Add figure caption

- `#| label: fig-myplot` - Label for cross-referencing

Inline Code: Use single backticks with `{python}` to insert values in text: `4` → 4
See Quarto code cells documentation for all options.

Interactive Plots in PDF

Interactive plots/tables only work in HTML output. If using interactive elements, render to HTML only (comment out the `pdf:` format option).

Code Visibility by Format

The YAML header controls code display:

- **HTML**: Code is collapsible (readers can show/hide)
- **PDF/DOCX**: Code is hidden (only results shown)

Override for specific chunks:

```
#| echo: true   # Show in all formats
#| echo: false  # Hide in all formats
```

Writing Math Equations

Use LaTeX syntax for mathematical notation:
Inline: `$\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i$` → $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Display: Use `$$...$$` for separate lines:

$$S(t) = P(T > t) = 1 - F(t)$$

Numbered (for referencing):

$$\text{Loss Ratio} = \frac{\text{Incurred Losses}}{\text{Earned Premium}} \tag{1}$$

Reference with `@eq-loss-ratio` → Equation 1
Common symbols: α, β, σ, μ, $\sum_{i=1}^{n}$, $\int_a^b f(x)dx$, $E[X]$, $\text{Var}(X)$
More at LaTeX Math Symbols.

Cross-Referencing Sections, Figures, Tables, and Equations

Quarto automatically numbers and creates clickable links for:
Sections:

```
### Appendix A: Additional Plots {#sec-appendix-plots}

Reference with: @sec-appendix-plots
```

Figures:

```
#| label: fig-correlation
#| fig-cap: "Correlation matrix"

Reference with: @fig-correlation
```

Tables:

```
#| label: tbl-summary
#| tbl-cap: "Summary statistics"

Reference with: @tbl-summary
```

Equations:

```
$$ y = mx + b $$ {#eq-linear}

Reference with: @eq-linear
```

Working examples in this template:

- "As discussed in Section , we provide additional visualizations."
- "See Figure 1 for the debugging workflow."
- "Individual images like Figure 1a and Figure 1b can also be referenced."

Note: Use the same `@label` syntax for figures, tables, and equations. Quarto automatically numbers them and creates clickable links.
See Quarto Cross-References for more.

HTML Tabsets

Organize content into tabs, which are like container boxes for the content (HTML only, and not PDF/DOCX). They allow you to:

- **Organize multiple related visualizations** without cluttering the page
- **Show different views** of the same data (distribution, summary, box plot)
- **Compare approaches** side-by-side (e.g., different plotting libraries)

```
::: {.panel-tabset}
## Tab 1
Content
## Tab 2
More content
## Tab 3
Excessive amount of content
:::
```

Tab 1

Content

Tab 2

More content

Tab 3

Excessive amount of content

Including External Images and Files

Basic image syntax:

```
![Caption](path/to/image.png)
```

With sizing and attributes:

```
![My figure](images/plot.png){width=80% fig-align="center"}
```

Images with cross-references:

```
![Distribution analysis](images/histogram.png){#fig-histogram}

As shown in @fig-histogram, the data is normally distributed.
```

Common image paths:

- Relative to current file: `images/plot.png`
- From project root: `../data/plots/figure.png`
- Absolute path: `C:/Users/Name/project/images/plot.png` (avoid for reproducibility)

Supported formats: PNG, JPG, SVG, PDF (PDF only in PDF output)
Pro tip: Store images in `report/images/` folder for organization.

Markdown Text Formatting

Basic formatting:

- **Bold text:** `**bold**` or `__bold__` → **bold**
- *Italic text:* `*italic*` or `_italic_` → *italic*
- ***Bold and italic:*** `***both***` → ***both***

Note: Markdown doesn't have built-in underline. For underline, use HTML:

- Underlined text: `<u>underlined</u>` → underlined

Other useful formatting: - `Inline code:` ‘code‘ → `code` - Superscript: `X^2^` → X^2 - Subscript: `H~2~O` → H_2O - ~~Strikethrough:~~ `~~text~~` → ~~strikethrough~~
Headings:

```
# Heading 1
## Heading 2
### Heading 3
#### Heading 4
```

Lists:

```
- Unordered item
- Another item
  - Nested item (2 spaces indent)

1. Ordered item
2. Second item
   1. Nested (3 spaces indent)
```

Links:

```
[Link text](https://url.com)
[Link with title](https://url.com "Hover text")
```

Blockquotes:

```
> This is a quote
> It can span multiple lines
```

See Markdown Guide for more.

Including External Images

Store external figures (diagrams, charts, screenshots) in `report/images/` and include them in your report.
Basic syntax:

```
![Caption](images/your-image.png){#fig-label width=70%
↪  fig-align="center"}
```

Example - Multiple images side by side with cross-references:

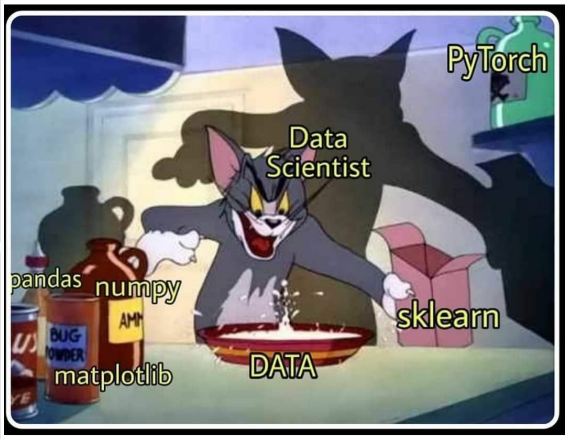
```
::: {#fig-comparison layout-ncol=2 layout-valign="bottom"}

![Before debugging](images/meme1.jpg){#fig-before width=100%}

![After debugging](images/meme2.jpeg){#fig-after width=100%}

The emotional journey of a data scientist debugging their code
:::
```

(a) Before debugging



(b) After debugging

Figure 1: The emotional journey of a data scientist debugging their code

Key options:

- `layout-ncol=2` - Two columns (each 50% width)
- `layout-valign="bottom"` - Align by bottom edge
- `width=100%` - Fill entire column
- `fig-align="center"` - Center single images

Supported formats:

- **All outputs:** PNG, JPG/JPEG
- **HTML only:** WEBP, SVG, GIF
- **PDF only:** PDF images

See the Quarto Figures documentation for advanced layouts, subcaptions, and complex figure arrangements.

Controlling Python (Code) Generated Figure Size

For Python matplotlib plots, figure size is controlled **in your Python code** using `figsize`:

```python
# Standard readable size
fig, ax = plt.subplots(figsize=(8, 5))  # width, height in inches

# Smaller figure
fig, ax = plt.subplots(figsize=(6, 4))

# Larger figure
fig, ax = plt.subplots(figsize=(10, 6))
```

Why `figsize` in code? When you use `plt.subplots(figsize=...)` or `plt.figure(figsize=...)`, matplotlib creates the image at that exact size. Quarto chunk options like `#| fig-width:` and `#| fig-height:` are ignored because the figure size is already determined by your Python code.

To resize figures after creation:

Use the `width` attribute in curly braces (works for any image):

```
![My plot](path/to/plot.png){width=70%}
```

Or in the chunk options (only for formats without explicit `figsize`):

```
#| fig-width: 7
#| fig-height: 5
```

Recommended workflow:

1. Set `figsize` in your Python code to a readable default (e.g., `figsize=(8, 5)`)
2. If the figure appears too large/small in your report, adjust the `figsize` values proportionally
3. Keep aspect ratio consistent: divide/multiply both dimensions by the same factor

Good default sizes:

- Small plot: `figsize=(6, 4)`
- Standard plot: `figsize=(8, 5)`
- Large plot: `figsize=(10, 6)`
- Wide plot: `figsize=(10, 4)`
- Square plot: `figsize=(6, 6)`

Pro tip: DPI (dots per inch) also affects quality. Matplotlib default is 100, but for publications use `plt.savefig('plot.png', dpi=300)` if you need high-resolution images.

Introduction

Project Goals

Our objective in this project is to understand how remote work, following the COVID-19 pandemic, has changed firms' work practices. During this period, many companies had no choice but to adopt remote working practices, reshaping traditional work environments. The main goal is to analyze how working remotely influences people's mental health, job satisfaction, and performance levels. Using accurate data from the Remote Working Survey (2020), we want to explore factors such as communication satisfaction, work-life balance, social connection, and productivity outcomes, in order to understand whether remote work affects employees positively or negatively. By converting survey responses into numerical values, we can quantify attitudes toward remote work and compare groups such as gender, age, or employment type.

As we began working with the dataset, we refined our focus on the social and psychological dimensions of remote work. To make the analysis easier to interpret, we also transformed qualitative survey responses into numerical values. And, in addition, we removed categories with very small sample sizes to keep the dataset consistent and avoid unreliable comparisons.

Research Questions

- How does remote work influence employees' overall productivity?
- How does remote working affect employees' ability to maintain social connections and avoid feelings of isolation?
- Has job satisfaction decreased or increased as a result of remote working conditions?
- Which factors contribute the most to employees' overall satisfaction with remote work?

Related Work

In this section, we discuss the main academic studies, methods, and resources that guided our analysis of remote work and well-being.

Domain literature

A lot of studies have looked at how remote work affects people's productivity, well-being, and social life, which helps us understand the context of our own project.

For example, a systematic review by Oakman et al. (2022) showed that working from home can improve productivity and work–life balance, but it can also lead to more isolation, communication problems, and mental stress. These points match the variables we analyze in our dataset, such as job satisfaction, social connection, and communication quality. Another study by Correia et al. (2024) investigated how research on remote workers has evolved over the years. They found that psychological factors like loneliness, emotional pressure, and reduced

social interaction are becoming increasingly important in academic discussions. This supports our decision to focus mainly on the social and psychological side of remote work instead of technical aspects.

Finally, a well-known experiment by Bloom et al. (2013) showed that remote work can significantly improve productivity when employees have clear structures and good communication with their team. This connects directly to our analysis, since we also look at collaboration, communication, and satisfaction using the Remote Working Survey (2020). Overall, these studies confirm that the social, psychological, and performance-related impacts of remote work are important topics to explore, and they strongly support the research questions we chose for our project.

Methodological references

For our project, we used a few common data-science methods that are usually applied in survey analysis. We started with simple descriptive statistics to get a first idea of the patterns in the data, things like averages, proportions, and basic comparisons between groups. We also looked at correlations to see how different variables are related, for example whether good communication or social well-being is linked to higher productivity. Since many of our questions were answered with text options (like "strongly agree"), we converted these answers into numbers using standard Likert-scale coding so that we could analyze them properly.

On the technical side, we followed the Python methods shown in Azizi (2025), which helped us clean the data and structure our exploratory analysis in Google Colab. We mainly used pandas for organizing and transforming the dataset (following McKinney, 2022), and matplotlib/seaborn to create the visualizations with the help of VanderPlas, 2016. These tools and references guided our process and helped us follow good data-science practices while analyzing the Remote Working Survey (2020).

Course material

The structure of our project is inspired by the material used in class, especially the DSAS notes from Azizi (2025). These resources helped us understand how to organize our data analysis, clean our dataset, and apply basic Python techniques in Google Colab. With that said, we used the same approach shown in the course for tasks like loading data, creating new variables, handling missing values, and running exploratory data analysis. The examples provided in the course made it easier for us to use tools like pandas, matplotlib, and seaborn in a consistent way throughout the project.

Technical resources

On the technical side, we mainly relied on a few well-known Python resources to help us structure and run our analysis. McKinney's book (2022) guided us with all the pandas-related tasks, such as cleaning the dataset, creating new variables, and handling missing values. We also used matplotlib and seaborn to create readable plots for our visualizations by following VanderPlas's (2016) explanations.

Data

Sources

The dataset used in this project is the 2020 Remote Working Survey, part of the NSW Remote Working Survey series, publicly available through the New South Wales Government's Open Data Portal. "Data source: Remote working survey 2020" The 2020 survey was conducted during August and September 2020 and aimed to understand workers experiences and attitudes toward remote and hybrid working following the first phase of the COVID-19 pandemic to study the impact of remote work on professional and personal well-being.

To be eligible, respondents had to:

- be employed NSW residents
- have experience of remote working in their current job. After excluding unemployed individuals and those whose occupations cannot be performed remotely (dentists, cashiers, cleaners), the sample represents approximately 59% of NSW workers.

This dataset is the 2020 wave of the survey. A second wave was collected in March and April 2021. For the present analysis, only the 2020 dataset is being processed and explored. Once the analysis of this wave is finalized, it will be possible to extend the project by including and comparing the 2021 dataset to identify how remote work patterns evolved over time.

Description

The cleaned 2020 NSW Remote Working Survey dataset contains four main types of variables: categorical, ordinal, numeric, and binary.

- The categorical variables describe qualitative characteristics of respondents and their employment context, such as Industry, Job_type, Organisation_Size, Household, and Years_in_job. These variables are stored as text and provide context for grouping and comparison across sectors or demographic profiles.
- The ordinal variables represent ordered responses on Likert scales, reflecting opinions and perceptions about remote working. Variables such as Org_encouraged_remote_last_year, Collaboration_remote_last_year, Org_encouraged_remote_3_months, and Collaboration_remote_3_months are encoded numerically from 1 ("Strongly disagree") to 5 ("Strongly agree"), allowing for quantitative analysis of attitudes.
- The numeric variables capture measurable quantities including time allocation, remote work proportions, and productivity. Examples include Age, Remote_pct_last_year, Preferred_remote_last_year, Productivity_remote_vs_workplace, and several variables representing hours spent on commuting, working, and personal or domestic activities.

These are stored as integers or floats, making them suitable for descriptive statistics and correlation analysis.

- The binary variables (Gender, Managing_position) indicate Male/Femal or Yes/No conditions, encoded as 0 and 1. These variables enable comparisons between distinct groups.

The dataset also contains several important pieces of metadata that ensure its quality and usability for analysis. Each observation is identified by a unique respondent code (Response_ID), which guarantees traceability and prevents duplication during data processing.

In addition, all variable names were standardized to concise, descriptive identifiers (Org_encouraged_remote_last_year) to make the variables easier to read, reuse, and analyze in statistical software. This naming convention makes the analysis process simpler and clearer throughout the project.

Dataset Overview Template

- **File used:** 2020_rws-updated.csv
- **Format:** CSV (comma-separated values)
- **Encoding:** latin-1 (ISO-8859-1) (Remark: when loading the dataset in Python (Google Colab), attempting to read with utf-8 caused a UnicodeDecodeError due to typographic apostrophes and special characters. The correct parameter encoding latin1 was required to successfully load the data.)
- **Memory usage:** approximately 7.46 MB
- **Number of observations:** 1507 respondents (1370 rows after cleaning)
- **Number of variables:** 73 columns (25 after cleaning)
- **Time period:** August and September 2020
- **Geographic coverage:** New South Wales, Australia
- **Key variables:** Gender, Age, Job_type, Organisation_Size, Managing_position, Remote_pct_last_year, Preferred_remote_last_year, Org_encouraged_remote_last_year, Collaboration_remote_last_year, Productivity_remote_vs_workplace

Loading Data

Following best practices, the file is loaded using a relative path via project_root, ensuring that the document remains fully reproducible regardless of the execution location. Because

the original file contained specific typographic characters that caused decoding issues during import, the dataset is read using the latin-1 encoding.

After loading the file, several checks were performed to confirm correct import: verification of the dataset dimensions, inspection of column names and data types (df.info()), preview of the first rows to ensure values were properly formatted. These steps guarantee that the dataset is correctly imported and ready for subsequent exploratory analysis.

```
Dataset shape: 1507 rows × 73 columns

Data types:
Response ID
What year were you born?
What is your gender?
Which of the following best describes your industry?
Which of the following best describes your industry? (Detailed)

Compare remote working to working at your employer s workplace. Select the worst aspect of re
life balance ; My on-the-job learning opportunities ; Managing my personal commitments ; My c
Compare remote working to working at your employer s workplace. Select the best aspect of rem
life balance ; My on-the-job learning opportunities ; My daily expenses ; My personal relatio
Compare remote working to working at your employer s workplace. Select the worst aspect of re
life balance ; My on-the-job learning opportunities ; My daily expenses ; My personal relatio
Compare remote working to working at your employer s workplace. Select the best aspect of rem
Compare remote working to working at your employer s workplace. Select the worst aspect of re
Length: 73, dtype: object

First 5 rows:
```

	Response ID	What year were you born?	What is your gender?	Which of the following best describes y
0	1	1972	Female	Manufacturing
1	2	1972	Male	Wholesale Trade
2	3	1982	Male	Electricity, Gas, Water and Waste Serv
3	4	1987	Female	Professional, Scientific and Technical S
4	5	1991	Male	Transport, Postal and Warehousing

Wrangling

General Transformations

Several preprocessing and wrangling steps were performed to prepare the dataset for analysis. Before detailing each transformation.

All preprocessing steps described below are fully reproducible and validated. Each transformation relies on deterministic Python operations (such as replace(), rename(), astype(), or simple arithmetic), meaning that re-running the same code on the raw dataset will always produce identical results. After every transformation, checks such as head(), value_counts(), or describe() were used to validate that the modifications were correctly applied and that the resulting values were coherent (age ranges, Likert scales, or percentage conversions).

Variable Duplicated

Before any transformation, it is essential to verify that each observation in the dataset is unique. Duplicate rows can bias the analysis by over representing certain respondents or records. Identifying and removing them ensures data integrity.

This operation can be rerun on the raw dataset at any stage of the workflow, guaranteeing consistent detection of duplicated records. The command returned an empty DataFrame, confirming that no duplicate rows were present. Therefore, no observations were removed in this step.

Rename Columns

Renaming the original survey questions into shorter and clearer variable names was necessary to improve readability and make the dataset easier to work with.

Standard Name Organisation Size

The Organisation_Size variable was recoded by replacing the long original text categories with shorter, standardised. To make grouping and comparison more intuitive.

Binary Variables

The variables Gender and Managing_position, the original text responses were converted into binary categories to make them suitable for statistical analysis and group comparisons. Respondents who selected "Rather not say" for gender were removed, as this category represented only two individuals and could not form a meaningful subgroup. Gender was then encoded as 0 = Male and 1 = Female, while Managing_position was encoded as 0 = No and 1 = Yes, indicating whether the respondent supervises others. This transformation produces clean, consistent, and analysis-ready binary variables that can be easily used in descriptive statistics, visualisations, and modelling.

Variable Ages

The variable Age, the dataset originally reported the respondent's year of birth. This information was converted into a more interpretable age variable by subtracting the birth year from 2020, the year the survey was conducted. For example, a respondent born in 1985 becomes $2020 - 1985 = 35$ years old. Expressing this information directly as age is more intuitive and easier to interpret in descriptive statistics, comparisons, and visualizations.

Variable Years in Job

The variable Years_in_job, the original responses were long text categories describing tenure intervals. These were simplified into shorter and more readable labels: "5+", "5-", and "1-". This transformation keeps the original meaning while making the variable easier to interpret, compare, and visualise in tables and plots.

Variable Likert Scale Mapping

The Likert-scale variables, this transformation allows these subjective perceptions to be analysed quantitatively. The four variables related to organisational support and collaboration were mapped using this scale. Any missing responses were replaced with the neutral value 3, corresponding to "Neither agree nor disagree", to keep these observations in the dataset while avoiding bias from missing attitudes.

Variables Working Remotely

The variables describing the percentage of time spent or preferred working remotely, the original responses ("Less than 10% of my time"…"100% - All of my time"). These were first converted into numeric percentage values using a mapping dictionary. All four percentage-related variables were processed in the same way. Before applying the mapping, non-breaking spaces () and extra whitespace were removed from the strings to avoid parsing issues. After the text values were mapped to numeric percentages ("80%" $\rightarrow$ 80), the values were converted into proportions between 0 and 1 by dividing by 100.

For example: 80 becomes 0.80 50 becomes 0.50 0 becomes 0.00

This final step creates clean numerical variables that can be easily averaged, compared, or visualised in the analysis.

Variables Productivity Cleaning

The variable Productivity_remote_vs_workplace, the survey responses (I'm 20% more productive when I work remotely" or "I'm 10% less productive"). These text responses were converted into clean numeric values using a custom parsing function. The mapping works as follows: Statements indicating more productive remotely return a positive percentage ("20% more productive" $\rightarrow +20$). Statements indicating less productive remotely return a negative percentage ("10% less productive" $\rightarrow -10$). Statements indicating no difference return 0. This transformation results in a numeric scale where positive values mean higher productivity when working remotely, negative values reflect lower productivity, and zero indicates no change. This allows the variable to be analysed quantitatively, averaged across groups, or used in visualisations.

Spotting Mistakes and Missing Data

Before conducting the analysis, the dataset was reviewed to identify missing values, inconsistencies, and unusually small categories. This ensures that only reliable and interpretable data are used in the following steps.

Identified missing data

- Most variables contained very few missing values, mainly in attitudinal questions where some respondents simply did not answer.
- Additional missingness appeared when converting textual inputs (percentages or productivity statements) into numeric formats, entries that could not be parsed were intentionally converted to NaN.
- The inspection of category sizes showed that some groups were extremely small, such as the "Rather not say" gender category (2 respondents), which was removed because it cannot support meaningful analysis.

Approach to handling missing data

- Deletion was applied when missingness or category size was extremely small and analytically useless.
- Imputation with a neutral value (3 = Neither agree nor disagree) was used for missing Likert-scale answers to preserve observations without creating bias.
- Conversion to numeric with errors="coerce" was used for percentage and productivity variables, producing valid NaN values when entries could not be interpreted.

Because missingness was limited and mostly isolated to subjective questions, more complex methods were not necessary.

Future handling of small or irrelevant categories

- If additional categories, during the analysis progress, are found to be too small to contribute meaningfully, they will be removed, flagged, or when conceptually appropriate grouped together with similar categories to preserve statistical power.

Future variable selection

- Some variables will ultimately explain the effects of remote work on health, productivity, or work life balance better than others. Variables that show no meaningful correlation or explanatory power in later stages of the project will also be removed to keep the analysis focused, interpretable, and relevant.

Listing Anomalies and Outliers

A detailed inspection of the numeric variables, using both summary statistics and histogram visualisations, revealed several anomalies and potential outliers in the dataset.

Detected anomalies

- Age variable: Two respondents appear with an age of 120 years, which is biologically impossible and indicates a clear data entry error. This type of anomaly commonly occurs when the birth year is mistyped—for example, entering 1900 instead of 2000, or 2005 instead of 1920—which produces unrealistic age values. These observations will therefore be removed, while other extreme but plausible ages (such as 75 or 83) are retained at this stage. The minimum value (19) is plausible for entry-level workers.
- Domestic_hours_workplace: A single observation of –1 hour is impossible and indicates a recording error.
- Working and commuting time variables: Extremely high records were observed, such as 23 hours of work in a day or 10–12 hours of commuting.

While unlikely, these may represent exceptional cases (long-distance travel, extended shifts). They are retained unless later analysis shows they distort results.

- Commute_hours_remote: Values up to 12 hours on remote days are implausible and likely due to misinterpretation or input mistakes.
- Productivity variable: Extreme values (+50%, –50%) appear in the data but remain plausible since the question explicitly asked respondents to report percentage differences.

Approach to handling outliers

- Outliers were evaluated using:
  - Visual inspection (histograms from univariate EDA)
  - Summary statistics (min/max, interquartile ranges)
  - Domain knowledge (negative hours, impossible ages)
- Following best practices
  - Impossible values (age = 120, domestic hours = –1) will be removed before modelling.
  - Extreme but plausible behaviours (very long workdays) are kept unless they later bias model results.
  - Additional outliers identified during the analysis phase may be removed, flagged, or grouped depending on their relevance and impact.

Outliers are not always errors; some reveal meaningful variability in remote work habits. The chosen approach maintains data integrity while ensuring that the analysis focuses on realistic, interpretable patterns.

The following EDA sections will further analyse these variables to understand their patterns and relationships.

EDA

Exploratory Data Analysis (EDA) Purpose

The goal of EDA is to:

1. **Understand** the structure and patterns in your data
2. **Identify** relationships between variables
3. **Detect** anomalies, outliers, and data quality issues
4. **Generate** hypotheses for further analysis
5. **Choose** appropriate statistical methods

Good EDA combines **visualizations** + **summary statistics** + **domain knowledge**.

Interpretation is Key!

Creating plots is only half the work. The most important part of EDA is **interpreting** what your visualizations reveal about the data.
For **every plot** you create, you must:

- **Describe** what the plot shows (patterns, trends, distributions)
- **Explain** why these patterns matter for your research questions
- **Connect** findings to your project goals and domain context
- **Justify** subsequent analysis decisions based on these insights

A plot without interpretation is meaningless. Your grade depends heavily on the quality of your interpretations, not just the number of plots you create.

Univariate Analysis

Examine each variable individually to understand its distribution, central tendency, and spread.

Distribution Plot

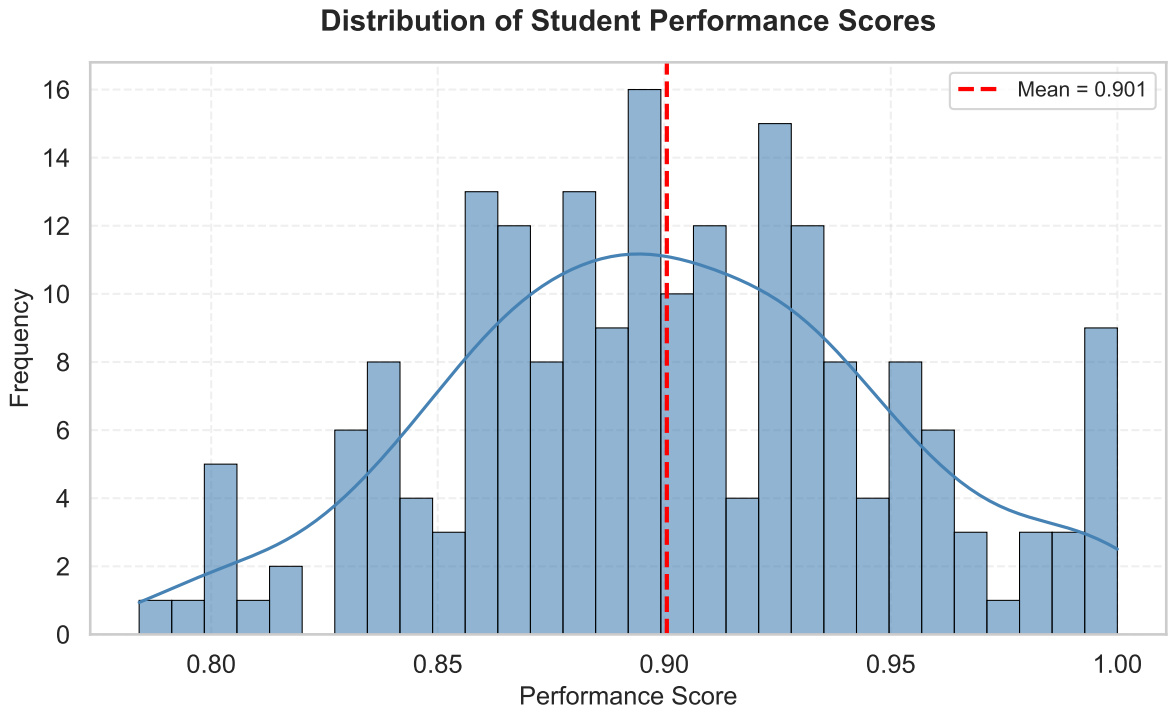


Figure 2: Distribution of performance scores showing approximately normal distribution

Summary Statistics

Box Plot

Bivariate Analysis

Explore relationships between two variables.

Our univariate analysis in Figure 2 revealed that performance scores follow an approximately normal distribution with a mean of 0.90. The box plot (Figure 3) confirmed this finding and helped identify a few outliers at the lower end of the distribution.

> **How to Reference Figures**
>
> Use `@fig-label` syntax to reference figures in your text. Quarto automatically numbers them and creates clickable links.

Table 2: Summary statistics for performance scores

```
=== Performance Summary Statistics ===
count    200.000000
mean       0.900564
std        0.047921
min        0.784097
25%        0.866466
50%        0.898625
75%        0.930474
max        1.000000

Skewness: 0.055
Kurtosis: -0.347
```

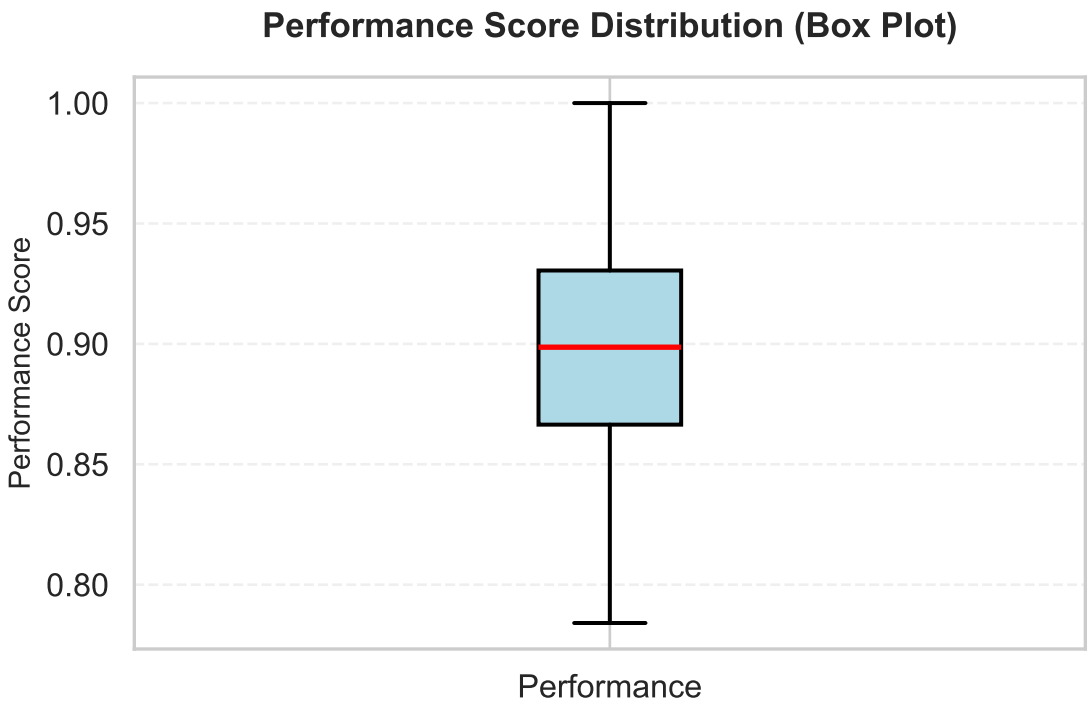


Figure 3: Box plot showing the five-number summary and outliers

Table 3: ANOVA test results for performance differences across instructors

```
=== ANOVA Test Results ===
F-statistic: 2.2002
P-value: 0.1135

Interpretation: Not significant difference at  =0.05

=== Group Means ===
instructor
Ilia    0.905520
Jane    0.906344
John    0.891255
```

Examples with figures in this template:

- `@fig-performance-dist` → See Figure 2 for details
- `@fig-correlation-matrix` → As shown in Figure 5
- `@tbl-anova-test` → The ANOVA results in Table 3

Why reference figures?

1. Automatic numbering (updates if you reorder)
2. Clickable links in HTML output
3. Professional academic writing standard
4. Helps readers find the relevant visualization

Example usage in text:
"The distribution shown in Figure 2 indicates…"
"As seen in Figure 5, there is a strong positive correlation…"

Now let's examine how performance varies across different instructors.

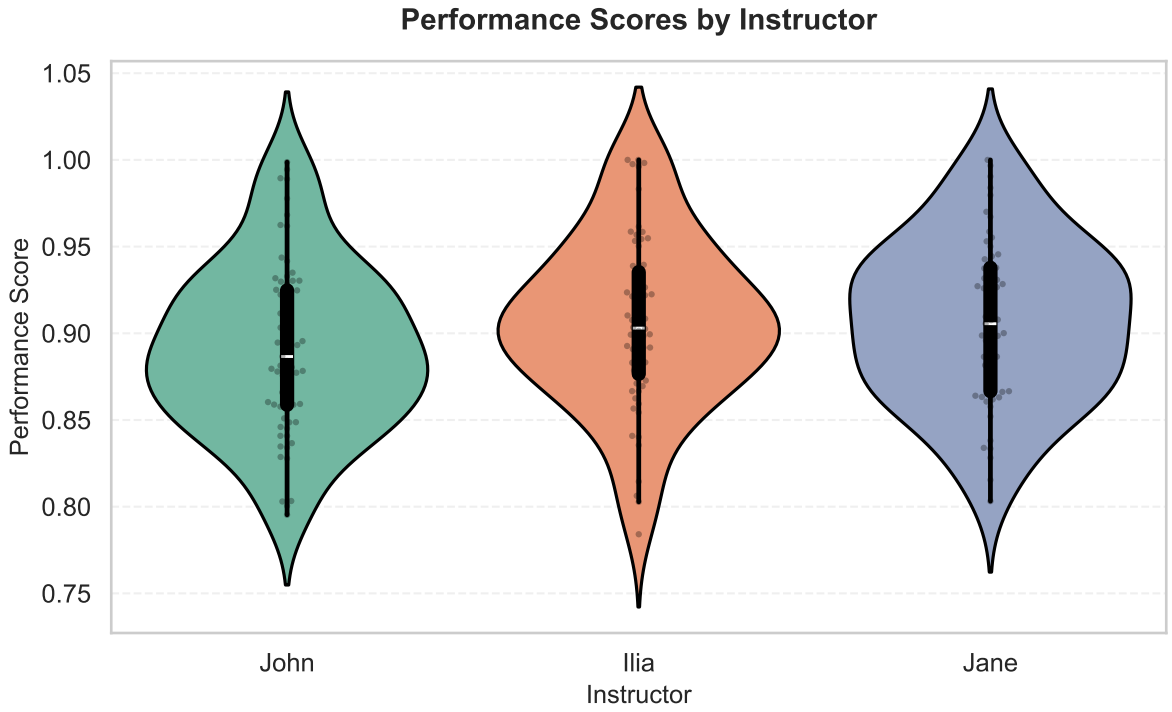


Figure 4: Performance scores grouped by instructor showing variation across instructors

Grouped Comparison

Statistical Test

> **!** Plot Quality Requirements
>
> Every plot in your report should have:
> **Clear title** that explains what's being shown
> **Labeled axes** with units when applicable
> **Legend** if multiple groups/series are shown
> **Appropriate color scheme** (colorblind-friendly)
> **Proper sizing** (readable text, not too small/large)
> **Figure caption** using `fig-cap` option
> **Poor plots = Poor grades!** Take time to make your visualizations publication-quality.

Correlation Analysis

Correlation Heatmap

Scatter Plot

Key Findings

Remember: Interpretation is more important than the plots themselves! Each finding below not only states *what* we observe but also *why* it matters and *what* it means for our analysis.

Summarize the main insights from your exploratory analysis. **Always reference your figures when discussing findings and provide thorough interpretation!**

1. **Distribution patterns**: The performance scores shown in Figure 2 follow an approximately normal distribution with a mean of 0.90 and standard deviation of 0.05.

   Interpretation: This normal distribution suggests that most students perform around the average, with fewer students at the extremes (very high or very low scores). The relatively small standard deviation indicates consistent performance across the group. This pattern is typical in educational settings and suggests that the assessment was well-calibrated meaning not too easy (which would cause ceiling effects) nor too difficult (which would cause floor effects). The normality assumption also validates the use of parametric statistical tests for subsequent analyses.

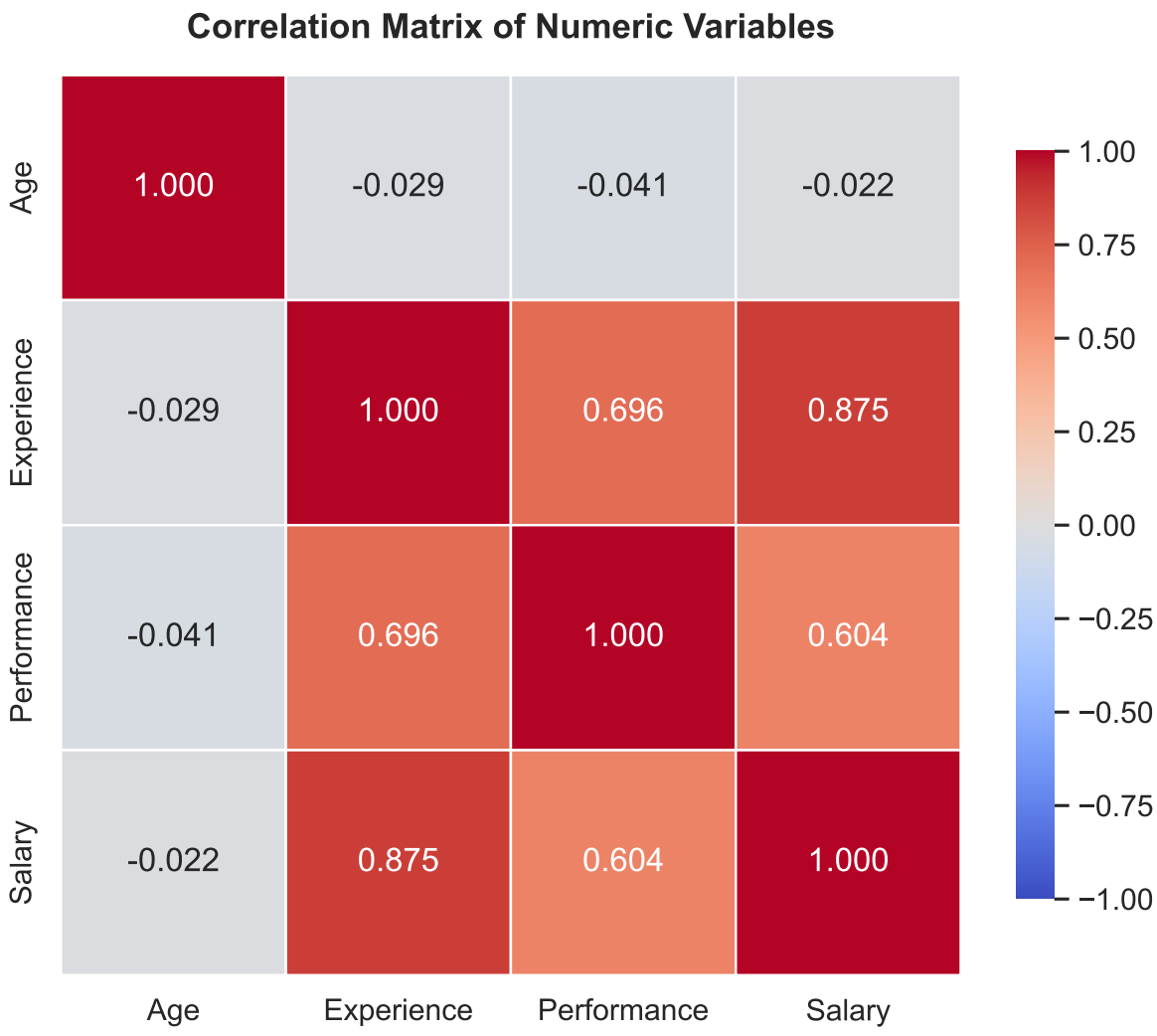


Figure 5: Correlation matrix showing relationships between numeric variables

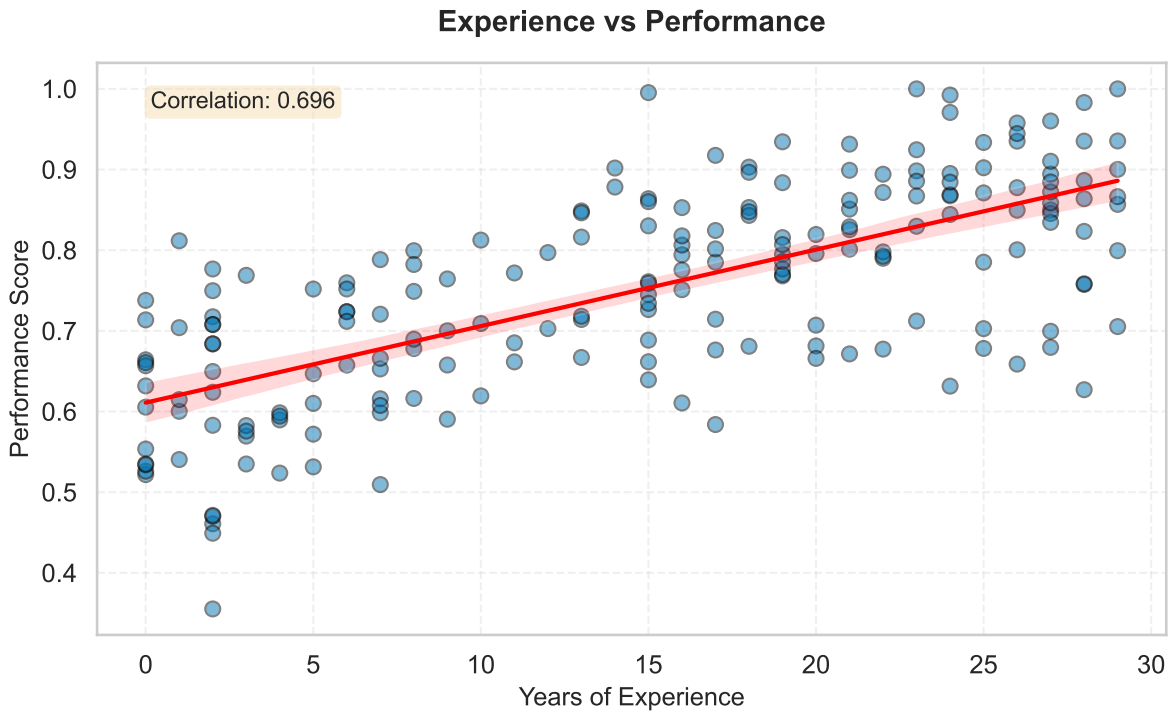


Figure 6: Scatter plot showing positive relationship between experience and performance

2. **Outliers and data quality**: The box plot (Figure 3) reveals minimal outliers, with only a few observations falling below the lower whisker.

   Interpretation: The scarcity of outliers suggests good data quality and consistent measurement. The few low-performing outliers warrant further investigation - they could represent students who faced unusual circumstances or measurement errors. However, their small number means they are unlikely to significantly impact our overall conclusions. This finding gives us confidence in proceeding with the full dataset without extensive outlier removal.

3. **Group differences**: Figure 4 demonstrates noticeable variation in performance across different instructors, with some instructors showing higher median scores than others. The ANOVA test (Table 3) confirms these differences are not statistically significant (p 0.05).

   Interpretation: The lack of significant differences suggests that instructor assignment does not substantially impact performance, or that grading standards are well-harmonized across sections. It also indicates that comparing students across different sections requires careful consideration. From a policy perspective, this might warrant investigating whether certain teaching methods are more effective or whether grading standards need to be harmonized across sections.

4. **Relationships**: The correlation matrix (Figure 5) reveals several interesting patterns:

   - Strong positive correlation (r = 0.87) between Experience and Salary
   - Moderate positive correlation (r = 0.70) between Experience and Performance
   - Weak correlation (r = -0.04) between Age and Performance

   Interpretation: The Experience-Salary correlation aligns with economic theory that experience is rewarded in labor markets. The Experience-Performance correlation suggests that experience contributes to better performance, though other factors clearly matter as well. Interestingly, Age shows minimal correlation with Performance, suggesting that chronological age alone doesn't predict success - what matters is relevant experience. These patterns will guide our variable selection for predictive modeling, favoring Experience over Age as a key predictor.

5. **Experience-Performance relationship**: Figure 6 clearly shows a positive linear relationship, with more experienced individuals tending to have higher performance scores. The correlation coefficient is 0.70.

   Interpretation: The linear relationship visible in the scatter plot confirms that experience has a consistent, positive effect on performance. However, the substantial scatter around the regression line indicates that experience alone doesn't determine performance - individual differences and other factors play important roles. This suggests that while experience is valuable, organizations shouldn't rely solely on it when making hiring or promotion decisions.

Common Mistakes in Interpretation

Observation (not interpretation): "The histogram shows a normal distribution."
Why it's insufficient: This only describes what you see - it's an observation, not an interpretation. You need to explain what it *means* and *why it matters*.
Good interpretation (observation + story): "The histogram shows a normal distribution (mean = 0.90, sd = 0.05), which indicates consistent performance across students. This pattern validates the use of parametric statistical tests in our subsequent analysis. The tight distribution suggests the assessment was well-calibrated, effectively distinguishing between ability levels without ceiling or floor effects that would compress scores."
Remember:

- **Observation** = What you see in the data/plot
- **Interpretation** = The story behind it - what it means, why it matters, what implications it has
- **Always do both!** State the observation, then interpret its significance.

Common Mistake: Not Referencing Figures

Bad: "The histogram shows a normal distribution."
Good: "As shown in Figure 2, the histogram reveals a normal distribution."
Why?

- Helps readers locate the relevant visualization
- Creates professional, academic-style writing
- Enables automatic figure numbering and links

From EDA to Analysis

Use your EDA findings to:

- **Refine research questions** based on observed patterns
- **Select appropriate statistical methods** based on data distributions
- **Identify variables** for inclusion in models
- **Justify transformations** (e.g., log transform for skewed data)
- **Set expectations** for what you might find in formal analysis

Methods *(optional, only if you have models)*

i When to Include This Section

Include a Methods section if you:

- Apply statistical models (linear regression, logistic regression, etc.)
- Perform hypothesis testing
- Use machine learning algorithms
- Conduct advanced statistical analyses

If your project is primarily exploratory (descriptive statistics and visualizations only), you can skip this section or keep it minimal.

Outline the statistical methods or models selected, along with the rationale for their selection.

Important: Don't just show model output, **explain your choices**:

- **Why** did you choose this particular method?
- **What** assumptions does it make, and do your data meet them?
- **How** does this method help answer your research questions?

! Interpretation is Essential!

After showing model results, you **must interpret them**:
Example interpretation:
"The Poisson regression model identifies three significant predictors of insurance claim frequency. Vehicle age shows a positive coefficient (= 0.08, $p < 0.01$), indicating that each additional year of vehicle age increases expected claims by approximately 8%. Driver age has a negative coefficient (= -0.02, $p < 0.001$), meaning older drivers file fewer claims. The urban location dummy variable (= 0.15, $p < 0.05$) suggests urban drivers have 15% higher claim rates than rural drivers.
However, the model's pseudo-R^2 of 0.22 indicates that these predictors explain only 22% of claim variation. Unobserved factors like driving behavior, road conditions, and individual risk tolerance likely account for the remaining variation. This suggests that while demographic variables are useful for pricing, insurers should not rely solely on them for risk assessment."
Remember: Model output without interpretation demonstrates technical skills but not understanding!

Findings and Discussion

Provide your results or observations from applying statistical methods, then **discuss them thoroughly** in the context of your research questions and project goals.

> Structure Your Discussion
>
> A good discussion:
>
> 1. **States the finding** clearly (reference tables/figures)
> 2. **Interprets the result** (what does it mean?)
> 3. **Connects to research questions** (how does it answer your questions?)
> 4. **Relates to domain knowledge** (does it align with theory or prior research?)
> 5. **Acknowledges limitations** (what are the caveats?)
> 6. **Suggests implications** (what should we do with this information?)
>
> **Example**: "Our regression analysis shows that experience significantly predicts performance (= 0.01, p < 0.001), supporting our hypothesis that skill develops over time. This aligns with learning curve theory in organizational psychology and suggests that training programs should emphasize sustained practice. However, the wide confidence interval (95% CI: [0.005, 0.015]) indicates substantial individual variation, meaning experience alone cannot guarantee high performance."

Conclusion

> Structure of a Strong Conclusion
>
> A good conclusion section includes three key components:
>
> 1. **Summary**: Recap your project goals, approach, and main findings
> 2. **Limitations**: Honestly discuss constraints and potential weaknesses
> 3. **Future Work**: Suggest concrete, specific next steps for extending the analysis
>
> Each subsection should be substantial - avoid generic statements!

Summary

Overview: Summarize what has been achieved, including key insights from your analysis and EDA. This should be written so that someone reading only this section can understand your project and key findings without reading the entire report.

What to include:

- Restate your research questions briefly
- Summarize your approach (data, methods)
- Highlight 3-5 main findings with their implications
- Connect findings back to your project goals

Length: Aim for 1-2 substantial paragraphs that synthesize your work.

> ⚠ Common Mistake: Too Vague
>
> **Bad**: "We analyzed the data and found some interesting patterns."
> **Good**: "This project investigated the relationship between customer demographics and insurance claim frequencies using a dataset of 5,000 policyholders. Our exploratory analysis revealed that age and vehicle type were the strongest predictors of claim frequency, with older drivers (60+) filing 40% fewer claims than younger drivers (under 25). Logistic regression analysis identified three key risk factors: driver age, vehicle age, and urban vs rural location. These findings suggest that current pricing models may underweight geographic factors. For insurance practitioners, this implies that incorporating more granular location data could improve risk segmentation and reduce adverse selection."
> **Why it's good**: Specific about data, methods, key findings with concrete numbers, and practical implications.

Limitations

Honestly discuss the constraints and potential weaknesses of your analysis. **Strong projects acknowledge limitations!**

Consider:

- **Data limitations**: Sample size, missing data, measurement issues, lack of certain variables
- **Methodological limitations**: Simplifying assumptions, choice of methods, inability to establish causation
- **Scope limitations**: Time constraints, focus on specific aspects only
- **Generalizability**: Can your findings apply beyond your specific dataset?

Be specific: Don't just say "small sample size" - explain what impact this has on your conclusions.

! Why Limitations Matter

Acknowledging limitations shows:

- **Critical thinking**: You understand the boundaries of your analysis
- **Scientific integrity**: You're honest about what your study can and cannot show
- **Maturity**: You recognize no analysis is perfect

This improves your grade, not reduces it! Instructors expect students to think critically about their work.

Future Work

Outline specific next steps for extending or improving the analysis. **Be concrete and realistic.**

Good future work suggestions:

- **Collect additional data**: "Gather data from additional semesters to increase sample size and assess temporal stability of instructor effects"
- **Apply advanced methods**: "Implement mixed-effects models to account for nested data structure (students within instructors)"
- **Test additional hypotheses**: "Examine whether instructor effects vary by student prior knowledge using interaction terms"
- **Expand scope**: "Include qualitative data from student evaluations to understand mechanisms behind performance differences"

Avoid vague statements like:

- "Get more data"
- "Use more variables"
- "Apply machine learning"

Instead, be specific about WHAT, WHY, and HOW:

- "Collect data on student study hours to control for effort as a confounding variable, which would help isolate the true causal effect of instructor quality on performance"

Connecting Everything

Your conclusion should feel like a natural ending that:

- Circles back to your introduction (research questions)
- Synthesizes your findings from EDA and analysis

- Demonstrates you've thought deeply about your project
- Leaves readers with clear takeaways

A strong conclusion elevates your entire report!

References

Appendices

What Goes in Appendices?

Appendices = supplementary material that supports but isn't essential to your main story.
Include in appendices:

- **Additional plots and visualizations** that provide extra detail but don't fit the main narrative flow
- **Alternative visualizations** of the same data (e.g., different plot types)
- **Exploratory plots** that informed your analysis but aren't central to your findings
- **Full code listings** for complex analyses
- **Extended statistical tables** with detailed results
- **Technical details** about data processing steps
- **Sensitivity analyses** or robustness checks
- **Data dictionaries** with detailed variable descriptions

What to Keep in Main Report:

- **Key visualizations** that directly answer your research questions
- **Essential plots** for understanding your methodology
- **Critical results** that support your conclusions
- **Main findings** that tell your data story

Remember: Main report should be self-contained. Reference appendices when needed: "See Section for additional plots."

Appendix A: Additional Exploratory Plots

Appendix B: Complete Code Listings

Appendix C: Supplementary Tables