Homework 2 - Exploratory Data Analysis

Mouad Medini & Lirim Zahiri

2024-12-26

Contents

1 Preparation

2 Task 1: Days with temperatures below 10°C at Newark Airport

```
# Exploring the weather dataset
glimpse(weather)
```

```
## Rows: 26,115
## Columns: 15
## $ origin     <chr> "EWR", "EWR", "EWR", "EWR", "EWR", "EWR", "EWR", "EWR", "EW~
## $ year       <int> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013,~
## $ month      <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
## $ day        <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
## $ hour       <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 13, 14, 15, 16, 17, 18, ~
## $ temp       <dbl> 39.02, 39.02, 39.02, 39.92, 39.02, 37.94, 39.02, 39.92, 39.~
```

```
## $ dewp       <dbl> 26.06, 26.96, 28.04, 28.04, 28.04, 28.04, 28.04, 28.04, 28.~
## $ humid      <dbl> 59.37, 61.63, 64.43, 62.21, 64.43, 67.21, 64.43, 62.21, 62.~
## $ wind_dir   <dbl> 270, 250, 240, 250, 260, 240, 240, 250, 260, 260, 260, 330,~
## $ wind_speed <dbl> 10.35702, 8.05546, 11.50780, 12.65858, 12.65858, 11.50780, ~
## $ wind_gust  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 20.~
## $ precip     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ pressure   <dbl> 1012.0, 1012.3, 1012.5, 1012.2, 1011.9, 1012.4, 1012.2, 101~
## $ visib      <dbl> 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10,~
## $ time_hour  <dttm> 2013-01-01 01:00:00, 2013-01-01 02:00:00, 2013-01-01 03:00~
```

```
?weather
```

```
## starte den http Server für die Hilfe fertig
```

```
# Filter and count for temperatures below 10°C (10° in Fahrenheit = 50) at Newark (EWR)
weather |>
  filter(origin == "EWR", temp < 50) |>
  group_by(month, date = as.Date(time_hour)) |>
  summarize(days_below_10 = n(), .groups = "drop") |>
  count(month, name = "days_below_10")
```

```
## # A tibble: 9 x 2
##   month days_below_10
##   <int>         <int>
## 1     1            32
## 2     2            29
## 3     3            32
## 4     4            22
## 5     5            12
## 6     9             3
## 7    10            12
## 8    11            28
## 9    12            28
```

(Slides used: EMPR_05_Data_Transformation_I_AS2024.pdf, EMPR_08a_EDA_AS2024.pdf)

3 Task 2: Analyze flights missing `dep_time`

```
# Flights with missing `dep_time` in january
flights |>
  filter(month == 1, is.na(dep_time))
```

```
## # A tibble: 521 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
##  1  2013     1     1       NA           1630        NA       NA           1815
##  2  2013     1     1       NA           1935        NA       NA           2240
##  3  2013     1     1       NA           1500        NA       NA           1825
##  4  2013     1     1       NA            600        NA       NA            901
```

```
##  5  2013     1     2       NA           1540        NA       NA           1747
##  6  2013     1     2       NA           1620        NA       NA           1746
##  7  2013     1     2       NA           1355        NA       NA           1459
##  8  2013     1     2       NA           1420        NA       NA           1644
##  9  2013     1     2       NA           1321        NA       NA           1536
## 10  2013     1     2       NA           1545        NA       NA           1910
## # i 511 more rows
## # i 11 more variables: arr_delay <dbl>, carrier <chr>, flight <int>,
## #   tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>,
## #   hour <dbl>, minute <dbl>, time_hour <dttm>
```

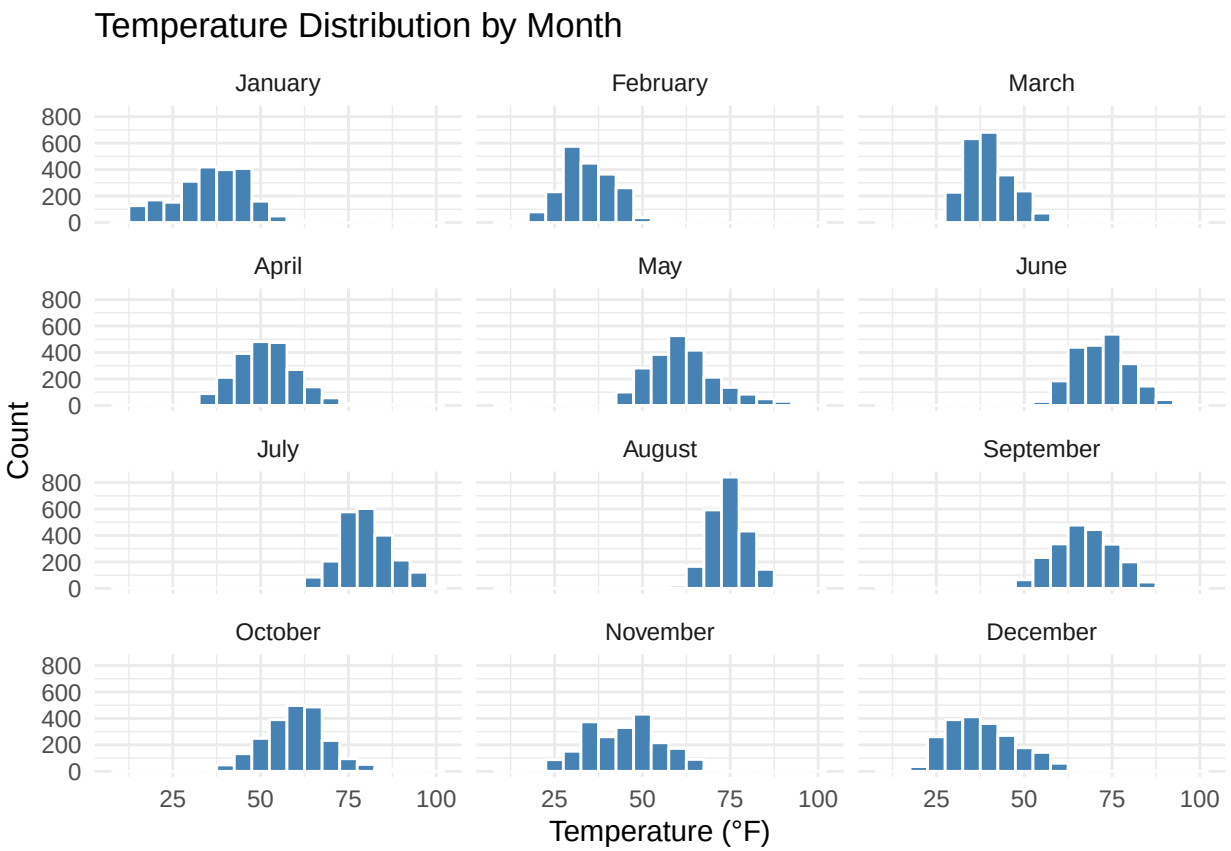
(Slides used: EMPR_05_Data_Transformation_I_AS2024.pdf)

4 Task 3: Visualizing temperatures

```r
# Select all variables except those between year and hour
# Add the month column from time_hour
weather_filtered <- weather |>
  select(-(year:hour)) |>
  mutate(month = month(time_hour))

# Visualization a
weather_filtered |>
  mutate(month = factor(month, labels = month.name)) |> # Convert month numbers to full names
  ggplot(aes(x = temp)) +
  geom_histogram(binwidth = 5, fill = "steelblue", color = "white") +
  facet_wrap(~ month, ncol = 3) +
  labs(title = "Temperature Distribution by Month", x = "Temperature (°F)", y = "Count") +
  theme_minimal()
```

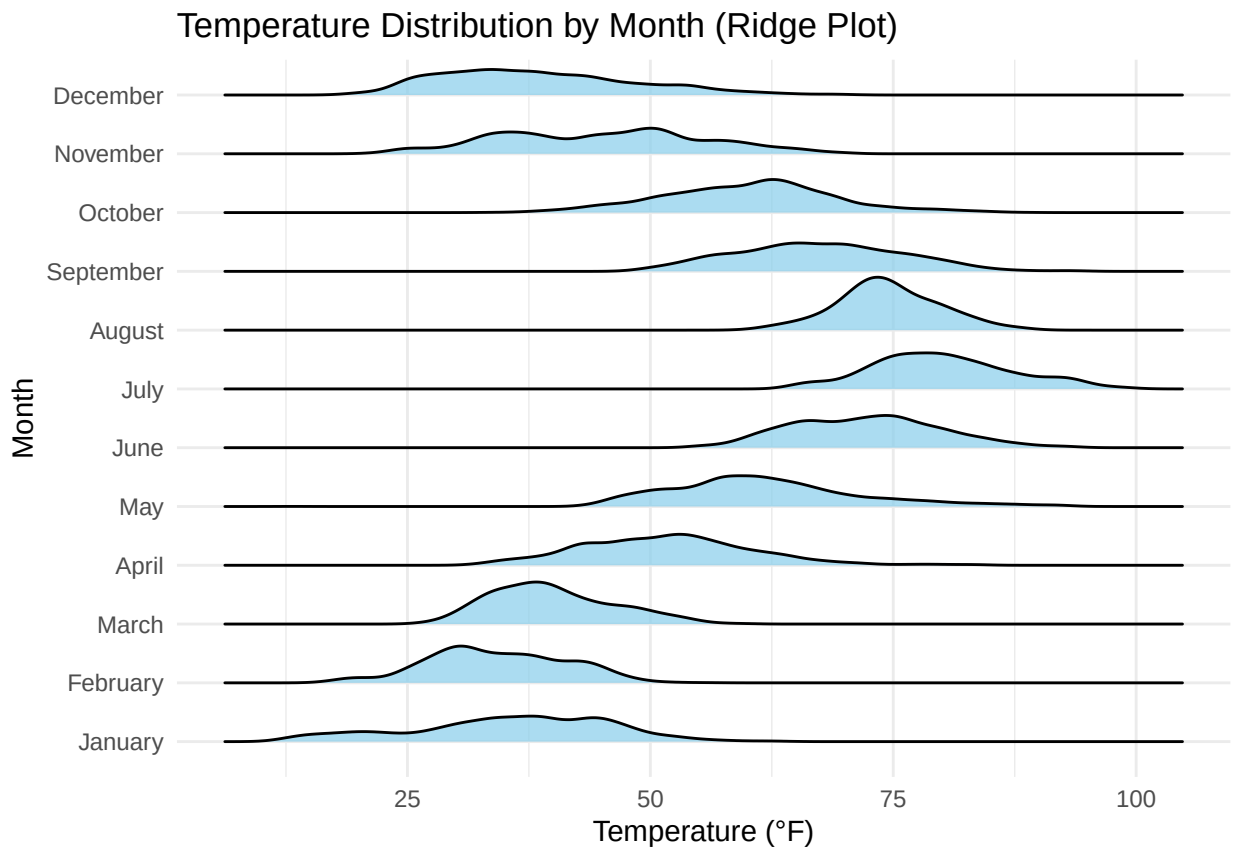
```
## Warning: Removed 1 row containing non-finite outside the scale range
## (‘stat_bin()‘).
```




```
# Ridge plot
weather_filtered |>
  mutate(month = factor(month, labels = month.name)) |> # Convert month numbers to full names
  ggplot(aes(x = temp, y = month)) +
  geom_density_ridges(scale = 0.9, fill = "skyblue", alpha = 0.7) + # Ridge plot
  labs(
    title = "Temperature Distribution by Month (Ridge Plot)",
    x = "Temperature (°F)",
    y = "Month"
  ) +
  theme_minimal()
```

```
## Picking joint bandwidth of 1.58
```

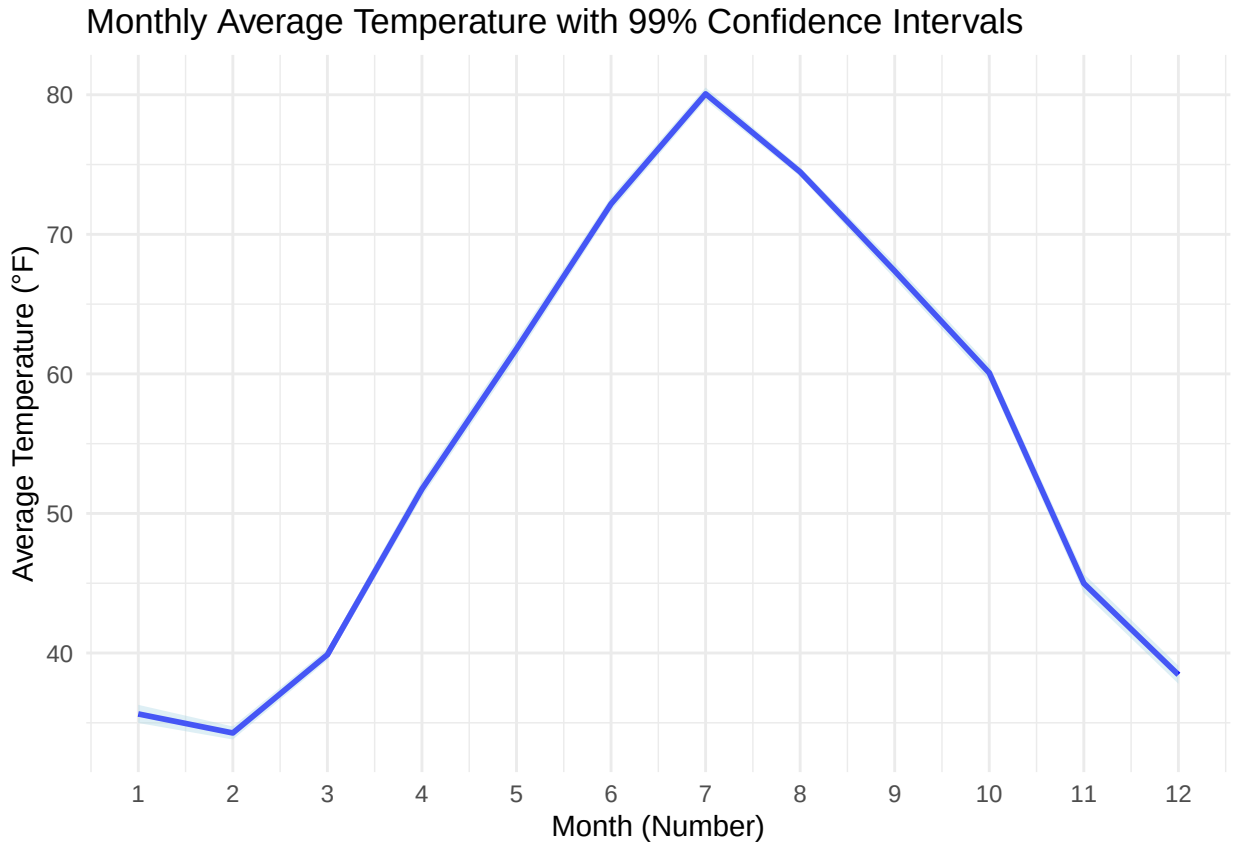
```
## Warning: Removed 1 row containing non-finite outside the scale range
## (`stat_density_ridges()`).
```




```r
# Aggregate data with weather_filtered
agg_data <- weather_filtered |>
  group_by(month) |>
  summarize(
    avg_temp = mean(temp, na.rm = TRUE),
    sd_temp = sd(temp, na.rm = TRUE),
    n = n()
  )

# Plot with explicit confidence interval calculation during visualization
ggplot(agg_data, aes(x = month, y = avg_temp)) +
  geom_line(size = 1, color = "blue") +                              # Line for average temperature
  geom_ribbon(
    aes(
      ymin = avg_temp - 3 * (sd_temp / sqrt(n)),                     # Lower 99% CI
      ymax = avg_temp + 3 * (sd_temp / sqrt(n))                      # Upper 99% CI
    ),
    fill = "lightblue", alpha = 0.4
  ) +
  scale_x_continuous(breaks = 1:12, labels = 1:12) +                 # Numeric month labels
  labs(
    title = "Monthly Average Temperature with 99% Confidence Intervals",
    x = "Month (Number)", y = "Average Temperature (°F)"
  ) +
  theme_minimal()
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



(Slides used: EMPR_03_Visualization1_AS2024.pdf, EMPR_07_Data_Tidying_AS2024,EMPR_05_Data_Transformatio EMPR_08a_EDA_AS2024.pdf)

5 Task 4: Data cleaning from `Data_HW2.xlsx`

```
# Import and cleaning the dataset
# Referenced from slides: EMPR_06_Import_Export_AS2024.pdf
data <- read_excel("Data_HW2.xlsx", sheet = "A very small sample")

data |> glimpse() |>
  summary()
```

```
## Rows: 18
## Columns: 10
## $ class           <chr> "2ab", "2ab", "2xyz", "2ab", "2xyz", "2ab", "2ab", "2a~
## $ gender          <chr> "Male", "Male", "Male", "Male", "Female", "Female", "M~
## $ ‘date of birth‘ <chr> "1992-08-28", "1991-02-14", "1995-11-29", "1997-09-04"~
```

```
## $ height          <chr> "178cm", "185cm", "179cm", "161cm", "163cm", "158cm", ~
## $ foot            <dbl> 26, 28, 45, 25, 24, 21, 27, 24, 26, 27, 43, 29, 24, 27~
## $ hair            <chr> "20", "6", "29", "4", "40", "35", "10", "35", "3", "10~
## $ `eye colour`    <chr> "Blau  Grau", "Grün", "Braun", "Braun", "Braun", "grün~
## $ `cash (CHF)`    <dbl> 25.6, NA, NA, 4000.0, 62.0, 40.0, 250.0, NA, 25.0, 25.~
## $ transport       <chr> "Bus", "Walk", "Bus", "Train", "Bus", "Train", "Train"~
## $ postcode        <dbl> 3074, 3007, 3037, 3172, 3004, 3257, 3270, 3072, 3032, ~
```

```
##     class              gender          date of birth         height
##  Length:18          Length:18          Length:18          Length:18
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##       foot          hair            eye colour          cash (CHF)
##  Min.   :21.0   Length:18          Length:18          Min.   :   1.00
##  1st Qu.:24.0   Class :character   Class :character   1st Qu.:  25.15
##  Median :26.0   Mode  :character   Mode  :character   Median :  45.00
##  Mean   :27.5                                         Mean   : 335.63
##  3rd Qu.:27.0                                         3rd Qu.:  57.55
##  Max.   :45.0                                         Max.   :4000.00
##                                                       NA's   :4
##   transport            postcode
##  Length:18          Min.   :2556
##  Class :character   1st Qu.:3009
##  Mode  :character   Median :3054
##                     Mean   :3171
##                     3rd Qu.:3255
##                     Max.   :3942
##
```

```r
data_clean <- data |>
  clean_names() |>
  mutate(
    # Convert class and gender to factors
    class = factor(class),
    gender = factor(gender),

    # Correct date_of_birth: Handle standard dates and Excel serial numbers
    date_of_birth = case_when(
      str_detect(date_of_birth, "^\\d{4}-\\d{2}-\\d{2}$") ~ as.Date(date_of_birth),  # Standard YYYY-MM
      str_detect(date_of_birth, "^\\d+$") ~ as.Date(as.numeric(date_of_birth), origin = "1899-12-30"),
      TRUE ~ NA_Date_  # Assign NA if unparseable
    ),

    # Correct height: Use parse_number to extract numeric values and handle cm/m conversion
    height = case_when(
      str_detect(height, "m") ~ parse_number(height) * 100,  # Convert "1,82m" to 182
      TRUE ~ parse_number(height)  # Parse "cm" values directly
    ) / 100,  # Fix incorrect scaling

    # Handle hair: Replace "Glatze" with "0" and convert to numeric
```

```
    hair = parse_number(if_else(hair == "Glatze", "0", hair)),

    # Standardize eye_colour, keeping combinations but removing "/"
    eye_colour = str_replace_all(
      eye_colour,
      regex("braun|brown", ignore_case = TRUE), "brown"
    ) |> str_replace_all(
      regex("blau|blue", ignore_case = TRUE), "blue"
    ) |> str_replace_all(
      regex("grün|green", ignore_case = TRUE), "green"
    ) |> str_replace_all(
      regex("grau|grey|gray", ignore_case = TRUE), "grey"
    ) |> str_replace_all(
      regex("schwarz|black", ignore_case = TRUE), "black"
    ) |> str_replace_all(
      "/", " "  # Replace "/" with a space
    ) |> str_squish()  # Remove extra spaces
  ) |>
  glimpse()
```

```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `date_of_birth = case_when(...)`.
## Caused by warning in `as.Date()`:
## ! NAs durch Umwandlung erzeugt
```

```
## Rows: 18
## Columns: 10
## $ class         <fct> 2ab, 2ab, 2xyz, 2ab, 2xyz, 2ab, 2ab, 2ab, 2xyz, 2xyz, 2x~
## $ gender        <fct> Male, Male, Male, Male, Female, Female, Male, Male, Male~
## $ date_of_birth <date> 1992-08-28, 1991-02-14, 1995-11-29, 1997-09-04, 1988-08~
## $ height        <dbl> 178, 185, 179, 161, 163, 158, 182, 182, 174, 181, 187, 1~
## $ foot          <dbl> 26, 28, 45, 25, 24, 21, 27, 24, 26, 27, 43, 29, 24, 27, ~
## $ hair          <dbl> 20, 6, 29, 4, 40, 35, 10, 35, 3, 10, 4, 0, 6, 6, 43, 28,~
## $ eye_colour    <chr> "blue grey", "green", "brown", "brown", "brown", "green ~
## $ cash_chf      <dbl> 25.6, NA, NA, 4000.0, 62.0, 40.0, 250.0, NA, 25.0, 25.0,~
## $ transport     <chr> "Bus", "Walk", "Bus", "Train", "Bus", "Train", "Train", ~
## $ postcode      <dbl> 3074, 3007, 3037, 3172, 3004, 3257, 3270, 3072, 3032, 30~
```

```
# Step 4: Save the cleaned data
write_rds(data_clean, "data_clean.rds")

# Step 5: Verify cleaned data
glimpse(data_clean)
```

```
## Rows: 18
## Columns: 10
## $ class         <fct> 2ab, 2ab, 2xyz, 2ab, 2xyz, 2ab, 2ab, 2ab, 2xyz, 2xyz, 2x~
## $ gender        <fct> Male, Male, Male, Male, Female, Female, Male, Male, Male~
## $ date_of_birth <date> 1992-08-28, 1991-02-14, 1995-11-29, 1997-09-04, 1988-08~
## $ height        <dbl> 178, 185, 179, 161, 163, 158, 182, 182, 174, 181, 187, 1~
## $ foot          <dbl> 26, 28, 45, 25, 24, 21, 27, 24, 26, 27, 43, 29, 24, 27, ~
## $ hair          <dbl> 20, 6, 29, 4, 40, 35, 10, 35, 3, 10, 4, 0, 6, 6, 43, 28,~
```

```
## $ eye_colour    <chr> "blue grey", "green", "brown", "brown", "brown", "green ~
## $ cash_chf      <dbl> 25.6, NA, NA, 4000.0, 62.0, 40.0, 250.0, NA, 25.0, 25.0,~
## $ transport     <chr> "Bus", "Walk", "Bus", "Train", "Bus", "Train", "Train", ~
## $ postcode      <dbl> 3074, 3007, 3037, 3172, 3004, 3257, 3270, 3072, 3032, 30~
```

(Slides used: EMPR_05_Data_Transformation_I_AS2024)

6 Task 5: Anscombe's Dataset

6.1 Part (a): Transform the Dataset into Tidy Format

```
# Transform Anscombe's dataset into tidy format

anscombe |> glimpse()
```

```
## Rows: 11
## Columns: 8
## $ x1 <dbl> 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
## $ x2 <dbl> 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
## $ x3 <dbl> 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
## $ x4 <dbl> 8, 8, 8, 8, 8, 8, 8, 19, 8, 8, 8
## $ y1 <dbl> 8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68
## $ y2 <dbl> 9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26, 4.74
## $ y3 <dbl> 7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73
## $ y4 <dbl> 6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56, 7.91, 6.89
```

```
tidy_anscombe <- anscombe |>
  pivot_longer(
    cols = everything(),
    names_to = c(".value", "sample"),
    names_pattern = "(.)(.)"
  ) |>
  glimpse()
```

```
## Rows: 44
## Columns: 3
## $ sample <chr> "1", "2", "3", "4", "1", "2", "3", "4", "1", "2", "3", "4", "1"~
## $ x      <dbl> 10, 10, 10, 8, 8, 8, 8, 8, 13, 13, 13, 8, 9, 9, 9, 8, 11, 11, 1~
## $ y      <dbl> 8.04, 9.14, 7.46, 6.58, 6.95, 8.14, 6.77, 5.76, 7.58, 8.74, 12.~
```

```
# Save the tidy dataset for future use
saveRDS(tidy_anscombe, file = "tidy_anscombe.rds")
```

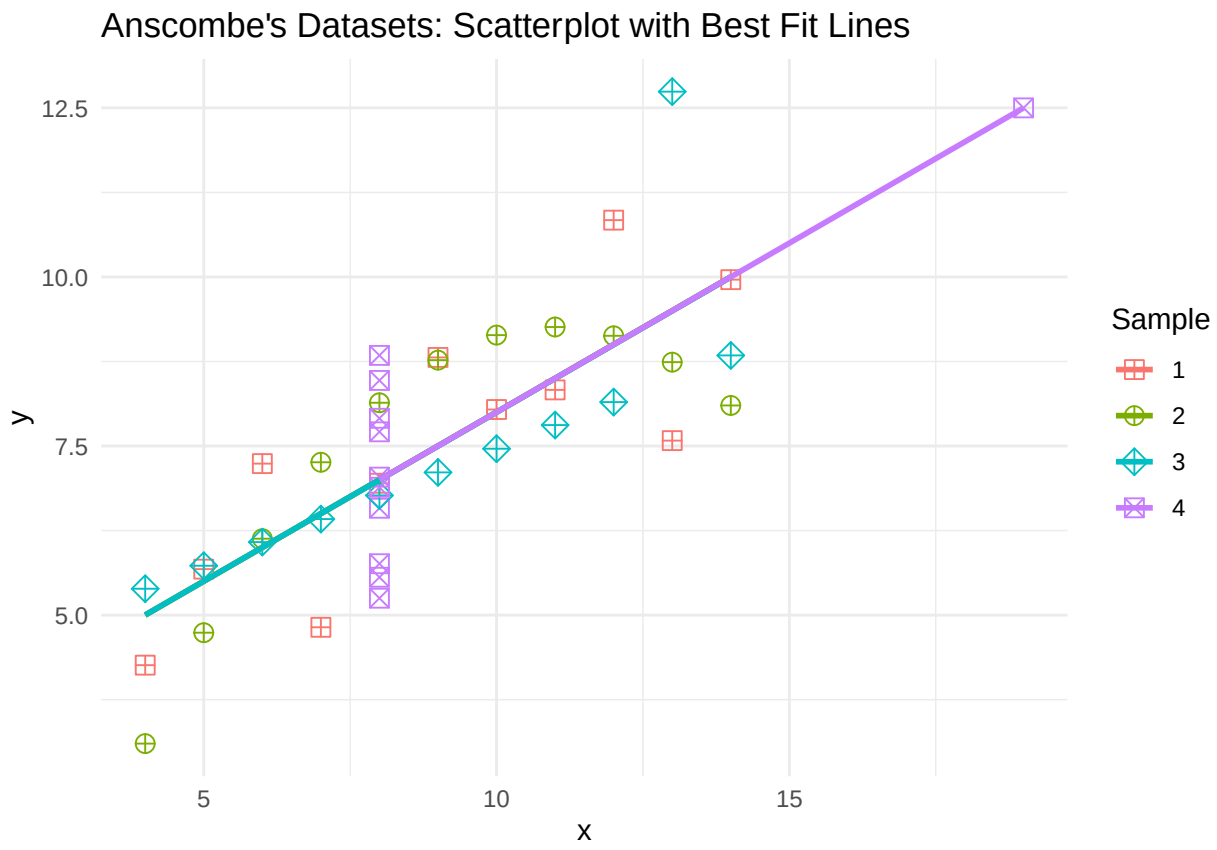
(Slides used: EMPR_07_Data_Tidying_AS2024)

6.2 Part (b): Scatterplot with Best Fit Lines

```r
# Load the tidy dataset
tidy_anscombe <- readRDS("tidy_anscombe.rds")

# Scatterplot with best fit lines
ggplot(tidy_anscombe, aes(x = x, y = y, color = sample, shape = sample)) +
  geom_point(size = 3) +
  geom_smooth(method = "lm", se = FALSE) +
  scale_shape_manual(values = c(12, 10, 9, 7)) +
  labs(
    title = "Anscombe's Datasets: Scatterplot with Best Fit Lines",
    x = "x",
    y = "y",
    color = "Sample",
    shape = "Sample"
  ) +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



(Slides used: EMPR_03_Visualization1_AS2024)

6.3 Part (c): Summarize the Tidy Dataset

```
# Summarize the tidy dataset and display using kableExtra in a single pipeline
tidy_anscombe |>
  group_by(sample) |>
  summarise(
    mean_x = mean(x),
    mean_y = mean(y),
    sd_x = sd(x),
    sd_y = sd(y),
    corr_xy = cor(x, y)
  ) |>
  kableExtra::kable(
    caption = "Summary of Anscombe's Dataset",
    col.names = c("Sample", "Mean x", "Mean y", "SD x", "SD y", "Correlation"),
    digits = 3
  )
```

Table 1: Summary of Anscombe's Dataset

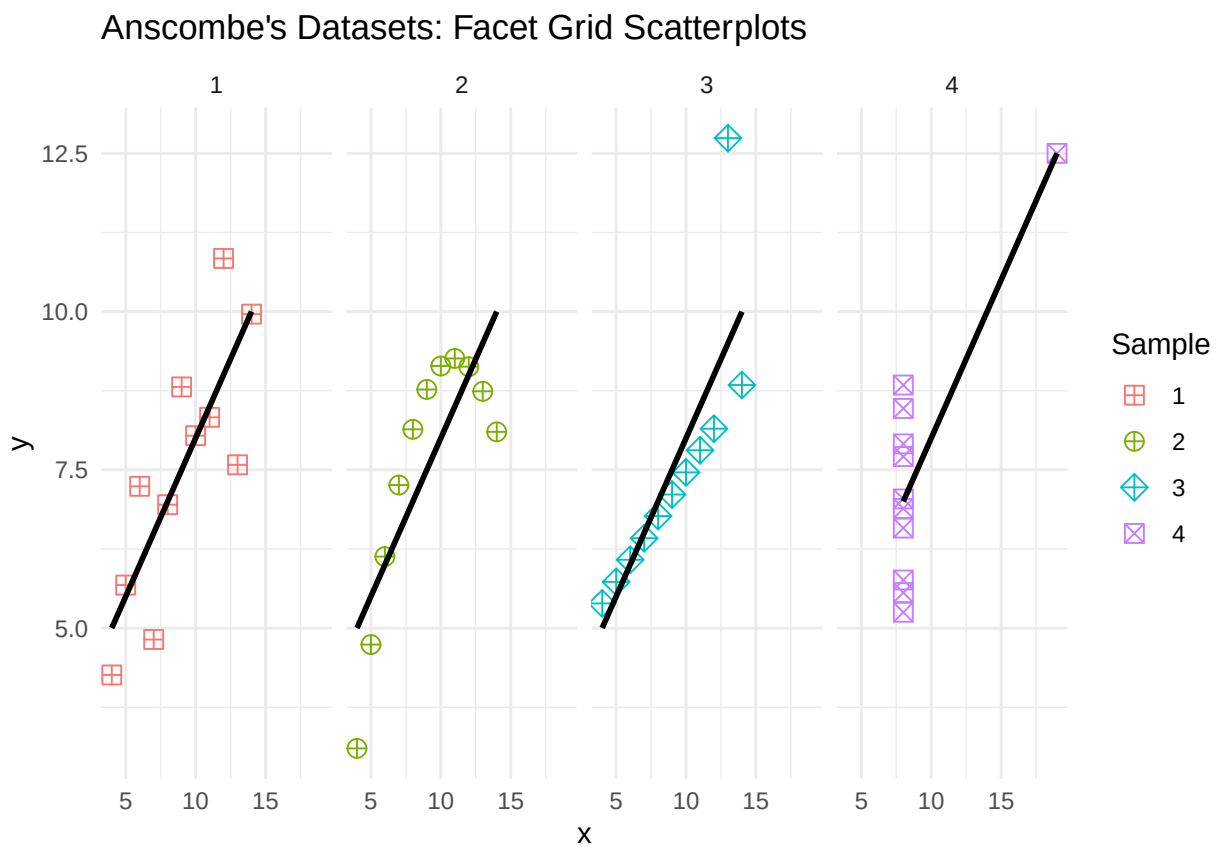
Sample	Mean x	Mean y	SD x	SD y	Correlation
1	9	7.501	3.317	2.032	0.816
2	9	7.501	3.317	2.032	0.816
3	9	7.500	3.317	2.030	0.816
4	9	7.501	3.317	2.031	0.817

(Slides used: EMPR_05_Data_Transformation_I_AS2024, EMPR_03_Visualization1_AS2024)

6.4 Part (d): Facet Grid Scatterplots

```
# Scatterplots with facet grid per sample
ggplot(tidy_anscombe, aes(x = x, y = y)) +
  geom_point(aes(color = sample, shape = sample), size = 3) +
  geom_smooth(method = "lm", se = FALSE, color = "black") +
  facet_grid(. ~ sample) +
  scale_shape_manual(values = c(12, 10, 9, 7)) +
  labs(
    title = "Anscombe's Datasets: Facet Grid Scatterplots",
    x = "x",
    y = "y",
    color = "Sample",
    shape = "Sample"
  ) +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



(Slides used: EMPR_03_Visualization1_AS2024)