Airline Pilot Retention Analysis

Mike Hickey

2025-05-16

Table of Contents

raw <- read\_csv(  
 here(params$data\_file), # data/retention.csv  
 show\_col\_types = FALSE  
)  
  
# Row 1 = headers; rows 2‑3 are survey text & metadata → drop them  
retention\_data <- raw %>%  
 slice(-(1:2)) %>% # remove rows 2 & 3  
 clean\_names() # snake\_case headers  
  
# Remove unwanted metadata columns  
ignore\_cols <- c(  
 "start\_date", "end\_date", "status", "progress", "duration\_in\_seconds",  
 "recorded\_date", "response\_id", "distribution\_channel",  
 "user\_language", "q\_recaptcha\_score"  
)  
retention\_data <- retention\_data %>% select(-all\_of(ignore\_cols))  
  
glimpse(retention\_data)

## Rows: 113  
## Columns: 52  
## $ finished <chr> "False", "False", "False", "False", "True", "True"…  
## $ informed\_consent <chr> "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "…  
## $ screener <chr> "Yes", "Yes", NA, NA, "Yes", "Yes", "Yes", "Yes", …  
## $ age <chr> NA, NA, NA, NA, "27", "28", "48", "34", "23", "42"…  
## $ gender <chr> NA, NA, NA, NA, "Male", "Male", "Female", "Male", …  
## $ nationality <chr> NA, NA, NA, NA, "United States of America", "Unite…  
## $ base\_state <chr> NA, NA, NA, NA, "California", "Ohio", "Montana", "…  
## $ position\_position <chr> NA, NA, NA, NA, "First Officer", "Captain", "Capta…  
## $ air\_carrier\_type <chr> NA, NA, NA, NA, "Low Cost Carrier", "Regional Airl…  
## $ experience <chr> NA, NA, NA, NA, "3", "3", "11", "4", "1", "2", "8"…  
## $ total\_flying\_hours <chr> NA, NA, NA, NA, "3000", "2800", "6800", "3600", "2…  
## $ military\_experience <chr> NA, NA, NA, NA, "Never served in the military", "N…  
## $ general\_1 <chr> NA, NA, NA, NA, "2", "1", "1", "1", "1", "1", "2",…  
## $ general\_2 <chr> NA, NA, NA, NA, "1", "4", "2", "2", "2", "6", "4",…  
## $ general\_3 <chr> NA, NA, NA, NA, "6", "2", "5", "3", "4", "3", "6",…  
## $ general\_4 <chr> NA, NA, NA, NA, "5", "5", "6", "6", "6", "2", "5",…  
## $ general\_5 <chr> NA, NA, NA, NA, "3", "3", "3", "4", "3", "4", "1",…  
## $ general\_6 <chr> NA, NA, NA, NA, "4", "6", "4", "5", "5", "5", "3",…  
## $ financial\_1 <chr> NA, NA, NA, NA, "1", "2", "1", "1", "1", "1", "1",…  
## $ financial\_2 <chr> NA, NA, NA, NA, "2", "5", "2", "4", "4", "3", "5",…  
## $ financial\_3 <chr> NA, NA, NA, NA, "4", "4", "3", "2", "2", "4", "2",…  
## $ financial\_4 <chr> NA, NA, NA, NA, "3", "3", "5", "6", "6", "5", "4",…  
## $ financial\_5 <chr> NA, NA, NA, NA, "5", "1", "4", "3", "3", "2", "3",…  
## $ financial\_6 <chr> NA, NA, NA, NA, "6", "6", "6", "5", "5", "6", "6",…  
## $ qo\_l\_1 <chr> NA, NA, NA, NA, "1", "3", "1", "3", "2", "1", "2",…  
## $ qo\_l\_2 <chr> NA, NA, NA, NA, "2", "6", "3", "4", "5", "5", "4",…  
## $ qo\_l\_3 <chr> NA, NA, NA, NA, "4", "4", "6", "6", "6", "6", "6",…  
## $ qo\_l\_4 <chr> NA, NA, NA, NA, "6", "2", "4", "1", "3", "2", "1",…  
## $ qo\_l\_5 <chr> NA, NA, NA, NA, "5", "5", "5", "5", "4", "3", "3",…  
## $ qo\_l\_6 <chr> NA, NA, NA, NA, "3", "1", "2", "2", "1", "4", "5",…  
## $ professional\_1 <chr> NA, NA, NA, NA, "1", "1", "1", "1", "2", "1", "1",…  
## $ professional\_2 <chr> NA, NA, NA, NA, "3", "4", "4", "2", "4", "2", "4",…  
## $ professional\_3 <chr> NA, NA, NA, NA, "4", "2", "3", "4", "3", "3", "5",…  
## $ professional\_4 <chr> NA, NA, NA, NA, "5", "5", "5", "3", "1", "5", "2",…  
## $ professional\_5 <chr> NA, NA, NA, NA, "2", "3", "2", "5", "5", "4", "3",…  
## $ recognition\_1 <chr> NA, NA, NA, NA, "2", "3", "3", "3", "2", "1", "4",…  
## $ recognition\_3 <chr> NA, NA, NA, NA, "1", "2", "2", "1", "3", "2", "1",…  
## $ recognition\_4 <chr> NA, NA, NA, NA, "3", "1", "1", "2", "1", "3", "3",…  
## $ recognition\_7 <chr> NA, NA, NA, NA, "4", "4", "4", "4", "4", "4", "2",…  
## $ schedule\_1 <chr> NA, NA, NA, NA, "5", "1", "1", "2", "5", "1", "1",…  
## $ schedule\_2 <chr> NA, NA, NA, NA, "3", "5", "4", "3", "3", "4", "5",…  
## $ schedule\_3 <chr> NA, NA, NA, NA, "1", "2", "5", "1", "2", "2", "2",…  
## $ schedule\_4 <chr> NA, NA, NA, NA, "4", "4", "2", "4", "1", "3", "4",…  
## $ schedule\_5 <chr> NA, NA, NA, NA, "2", "3", "3", "5", "4", "5", "3",…  
## $ operational\_1 <chr> NA, NA, NA, NA, "1", "1", "1", "3", "1", "3", "3",…  
## $ operational\_2 <chr> NA, NA, NA, NA, "2", "2", "4", "5", "2", "1", "1",…  
## $ operational\_3 <chr> NA, NA, NA, NA, "3", "4", "2", "1", "3", "2", "2",…  
## $ operational\_4 <chr> NA, NA, NA, NA, "5", "5", "5", "2", "5", "5", "5",…  
## $ operational\_5 <chr> NA, NA, NA, NA, "4", "3", "3", "4", "4", "4", "4",…  
## $ leaving <chr> NA, NA, NA, NA, "Maybe", "Yes", "Yes", "Yes", "Yes…  
## $ when <chr> NA, NA, NA, NA, "1-2 years", "1-2 years", "Within …  
## $ why <chr> NA, NA, NA, NA, "Bases in lower cost of living are…

total\_records <- nrow(retention\_data)  
  
cleaned <- retention\_data %>%  
 mutate(  
 finished = toupper(as.character(finished)),  
 informed\_consent = toupper(as.character(informed\_consent)),  
 screener = toupper(as.character(screener))  
 ) %>%  
 mutate(  
 exclusion\_reason = case\_when(  
 informed\_consent != "YES" ~ "did\_not\_consent",  
 finished != "TRUE" ~ "did\_not\_finish",  
 screener != "YES" ~ "not\_air\_carrier",  
 TRUE ~ "included"  
 )  
 )  
  
exclusion\_table <- cleaned %>%  
 count(exclusion\_reason, name = "n")  
  
knitr::kable(exclusion\_table, caption = "Why Records Were Excluded")

Why Records Were Excluded

| exclusion\_reason | n |
| --- | --- |
| did\_not\_finish | 35 |
| included | 76 |
| not\_air\_carrier | 2 |

included\_data <- cleaned %>%  
 filter(exclusion\_reason == "included") %>%  
 select(-exclusion\_reason) # drop helper column  
  
cat("\*\*Total records in file:\*\*", total\_records,  
 "\\n\\n\*\*Records kept for analysis:\*\*", nrow(included\_data),  
 "\\n\\n\*\*Records excluded:\*\*", total\_records - nrow(included\_data))

## \*\*Total records in file:\*\* 113 \n\n\*\*Records kept for analysis:\*\* 76 \n\n\*\*Records excluded:\*\* 37

write\_csv(included\_data, here("data/retention\_cleaned.csv"))

Preview of Cleaned Data

| finished | informed\_consent | screener | age | gender | nationality | base\_state | position\_position | air\_carrier\_type | experience | total\_flying\_hours | military\_experience | general\_1 | general\_2 | general\_3 | general\_4 | general\_5 | general\_6 | financial\_1 | financial\_2 | financial\_3 | financial\_4 | financial\_5 | financial\_6 | qo\_l\_1 | qo\_l\_2 | qo\_l\_3 | qo\_l\_4 | qo\_l\_5 | qo\_l\_6 | professional\_1 | professional\_2 | professional\_3 | professional\_4 | professional\_5 | recognition\_1 | recognition\_3 | recognition\_4 | recognition\_7 | schedule\_1 | schedule\_2 | schedule\_3 | schedule\_4 | schedule\_5 | operational\_1 | operational\_2 | operational\_3 | operational\_4 | operational\_5 | leaving | when | why |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| TRUE | YES | YES | 27 | Male | United States of America | California | First Officer | Low Cost Carrier | 3 | 3000 | Never served in the military | 2 | 1 | 6 | 5 | 3 | 4 | 1 | 2 | 4 | 3 | 5 | 6 | 1 | 2 | 4 | 6 | 5 | 3 | 1 | 3 | 4 | 5 | 2 | 2 | 1 | 3 | 4 | 5 | 3 | 1 | 4 | 2 | 1 | 2 | 3 | 5 | 4 | Maybe | 1-2 years | Bases in lower cost of living areas. |
| TRUE | YES | YES | 28 | Male | United States of America | Ohio | Captain | Regional Airline | 3 | 2800 | Never served in the military | 1 | 4 | 2 | 5 | 3 | 6 | 2 | 5 | 4 | 3 | 1 | 6 | 3 | 6 | 4 | 2 | 5 | 1 | 1 | 4 | 2 | 5 | 3 | 3 | 2 | 1 | 4 | 1 | 5 | 2 | 4 | 3 | 1 | 2 | 4 | 5 | 3 | Yes | 1-2 years | For a better airline that has better opportunities |
| TRUE | YES | YES | 48 | Female | United States of America | Montana | Captain | Freight/Cargo Carrier | 11 | 6800 | Never served in the military | 1 | 2 | 5 | 6 | 3 | 4 | 1 | 2 | 3 | 5 | 4 | 6 | 1 | 3 | 6 | 4 | 5 | 2 | 1 | 4 | 3 | 5 | 2 | 3 | 2 | 1 | 4 | 1 | 4 | 5 | 2 | 3 | 1 | 4 | 2 | 5 | 3 | Yes | Within 1 year | Unfair, unsafe, and fatiguing work rules/schedules. Company is not negotiating in good faith. |
| TRUE | YES | YES | 34 | Male | United States of America | Pennsylvania | Captain | Regional Airline | 4 | 3600 | Previously served in the military | 1 | 2 | 3 | 6 | 4 | 5 | 1 | 4 | 2 | 6 | 3 | 5 | 3 | 4 | 6 | 1 | 5 | 2 | 1 | 2 | 4 | 3 | 5 | 3 | 1 | 2 | 4 | 2 | 3 | 1 | 4 | 5 | 3 | 5 | 1 | 2 | 4 | Yes | 1-2 years | More pay, better 401k, better reserve schedule, higher seniority for non-reving, better ability to manipulate my scheudle. |
| TRUE | YES | YES | 23 | Male | United States of America | Ohio | First Officer | Regional Airline | 1 | 2400 | Never served in the military | 1 | 2 | 4 | 6 | 3 | 5 | 1 | 4 | 2 | 6 | 3 | 5 | 2 | 5 | 6 | 3 | 4 | 1 | 2 | 4 | 3 | 1 | 5 | 2 | 3 | 1 | 4 | 5 | 3 | 2 | 1 | 4 | 1 | 2 | 3 | 5 | 4 | Yes | 1-2 years | Transition to mainline, for better pay, and benefits |
| TRUE | YES | YES | 42 | Male | United States of America | Illinois | First Officer | Regional Airline | 2 | 2500 | Never served in the military | 1 | 6 | 3 | 2 | 4 | 5 | 1 | 3 | 4 | 5 | 2 | 6 | 1 | 5 | 6 | 2 | 3 | 4 | 1 | 2 | 3 | 5 | 4 | 1 | 2 | 3 | 4 | 1 | 4 | 2 | 3 | 5 | 3 | 1 | 2 | 5 | 4 | Maybe | 1-2 years | Fly larger aircraft that pays better |

library(DescTools) # for Skew() and Kurt()  
  
# ---- 1. make sure age is numeric ---------------------------------------  
included\_data <- included\_data %>%  
 mutate(age = as.numeric(age))  
  
# ---- 2. summary table ---------------------------------------------------  
age\_summary <- included\_data %>%  
 summarise(  
 n = sum(!is.na(age)),  
 min\_age = min(age, na.rm = TRUE),  
 q1\_age = quantile(age, 0.25, na.rm = TRUE),  
 median\_age = median(age, na.rm = TRUE),  
 mean\_age = mean(age, na.rm = TRUE),  
 q3\_age = quantile(age, 0.75, na.rm = TRUE),  
 max\_age = max(age, na.rm = TRUE),  
 sd\_age = sd(age, na.rm = TRUE),  
 iqr\_age = IQR(age, na.rm = TRUE),  
 skew\_age = Skew(age, na.rm = TRUE),  
 kurt\_age = Kurt(age, na.rm = TRUE)  
 )  
  
knitr::kable(age\_summary, digits = 2,  
 caption = "Descriptive Statistics for Age")

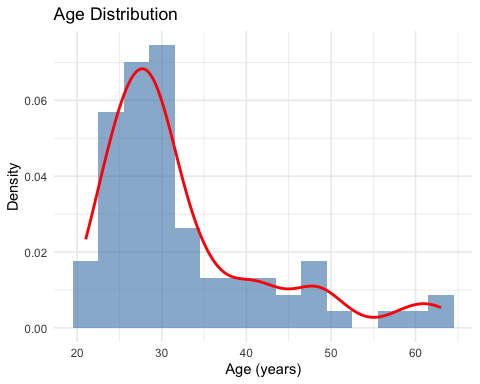
Descriptive Statistics for Age

| n | min\_age | q1\_age | median\_age | mean\_age | q3\_age | max\_age | sd\_age | iqr\_age | skew\_age | kurt\_age |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 76 | 21 | 26 | 29 | 32.41 | 35 | 63 | 9.91 | 9 | 1.46 | 1.48 |

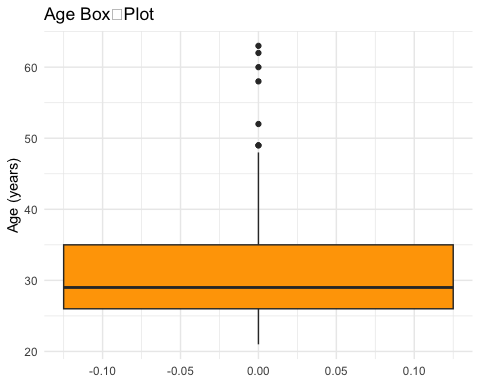
# ---- 3. normality test --------------------------------------------------  
if (age\_summary$n <= 5000) {  
 shapiro\_p <- shapiro.test(included\_data$age)$p.value  
 cat("\nShapiro–Wilk normality test p‑value:", round(shapiro\_p, 4), "\n")  
}

##   
## Shapiro–Wilk normality test p‑value: 0

# ---- 4. histogram + density --------------------------------------------  
ggplot(included\_data, aes(x = age)) +  
 geom\_histogram(aes(y = ..density..), bins = 15,  
 fill = "steelblue", alpha = 0.6) +  
 geom\_density(color = "red", linewidth = 1) +  
 labs(title = "Age Distribution",  
 x = "Age (years)", y = "Density") +  
 theme\_minimal()



# ---- 5. box‑plot for outliers ------------------------------------------  
ggplot(included\_data, aes(y = age)) +  
 geom\_boxplot(fill = "orange", width = 0.25) +  
 labs(title = "Age Box‑Plot", y = "Age (years)") +  
 theme\_minimal()



library(DescTools) # for Skew() and Kurt()  
  
# 1 Ensure the variable is numeric  
included\_data <- included\_data %>%  
 mutate(experience = as.numeric(experience))  
  
# 2 Summary table  
experience\_summary <- included\_data %>%  
 summarise(  
 n = sum(!is.na(experience)),  
 min\_exp = min(experience, na.rm = TRUE),  
 q1\_exp = quantile(experience, 0.25, na.rm = TRUE),  
 median\_exp = median(experience, na.rm = TRUE),  
 mean\_exp = mean(experience, na.rm = TRUE),  
 q3\_exp = quantile(experience, 0.75, na.rm = TRUE),  
 max\_exp = max(experience, na.rm = TRUE),  
 sd\_exp = sd(experience, na.rm = TRUE),  
 iqr\_exp = IQR(experience, na.rm = TRUE),  
 skew\_exp = Skew(experience, na.rm = TRUE),  
 kurt\_exp = Kurt(experience, na.rm = TRUE)  
 )  
  
knitr::kable(experience\_summary, digits = 2,  
 caption = "Descriptive Statistics for Years of Experience")

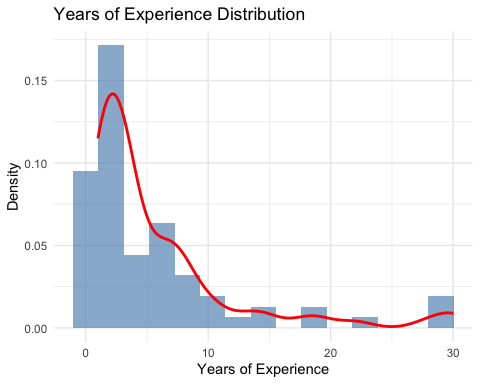
Descriptive Statistics for Years of Experience

| n | min\_exp | q1\_exp | median\_exp | mean\_exp | q3\_exp | max\_exp | sd\_exp | iqr\_exp | skew\_exp | kurt\_exp |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 76 | 1 | 2 | 3 | 5.78 | 7 | 30 | 6.55 | 5 | 2.21 | 4.74 |

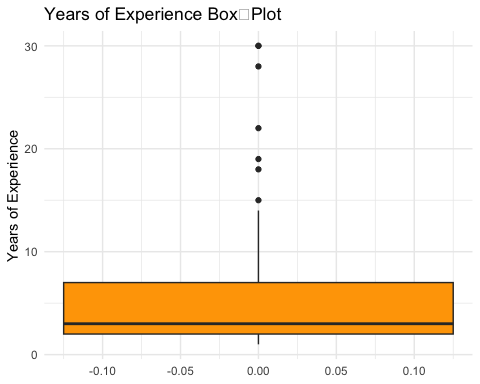
# 3 Normality test (Shapiro–Wilk)  
if (experience\_summary$n <= 5000) {  
 shapiro\_p <- shapiro.test(included\_data$experience)$p.value  
 cat("\nShapiro–Wilk normality test p‑value:",  
 format(shapiro\_p, digits = 4), "\n")  
}

##   
## Shapiro–Wilk normality test p‑value: 3.131e-11

# 4 Histogram + density curve  
ggplot(included\_data, aes(x = experience)) +  
 geom\_histogram(aes(y = ..density..), bins = 15,  
 fill = "steelblue", alpha = 0.6) +  
 geom\_density(color = "red", linewidth = 1) +  
 labs(title = "Years of Experience Distribution",  
 x = "Years of Experience", y = "Density") +  
 theme\_minimal()



# 5 Box‑plot  
ggplot(included\_data, aes(y = experience)) +  
 geom\_boxplot(fill = "orange", width = 0.25) +  
 labs(title = "Years of Experience Box‑Plot",  
 y = "Years of Experience") +  
 theme\_minimal()



library(DescTools) # Skew() and Kurt()  
  
# 1 Ensure the variable is numeric ---------------------------------------  
included\_data <- included\_data %>%  
 mutate(total\_flying\_hours = as.numeric(total\_flying\_hours))  
  
# 2 Summary statistics ----------------------------------------------------  
hours\_summary <- included\_data %>%  
 summarise(  
 n = sum(!is.na(total\_flying\_hours)),  
 min\_hours = min(total\_flying\_hours, na.rm = TRUE),  
 q1\_hours = quantile(total\_flying\_hours, 0.25, na.rm = TRUE),  
 median\_hours = median(total\_flying\_hours, na.rm = TRUE),  
 mean\_hours = mean(total\_flying\_hours, na.rm = TRUE),  
 q3\_hours = quantile(total\_flying\_hours, 0.75, na.rm = TRUE),  
 max\_hours = max(total\_flying\_hours, na.rm = TRUE),  
 sd\_hours = sd(total\_flying\_hours, na.rm = TRUE),  
 iqr\_hours = IQR(total\_flying\_hours, na.rm = TRUE),  
 skew\_hours = Skew(total\_flying\_hours, na.rm = TRUE),  
 kurt\_hours = Kurt(total\_flying\_hours, na.rm = TRUE)  
 )  
  
knitr::kable(hours\_summary, digits = 2,  
 caption = "Descriptive Statistics for Total Flying Hours")

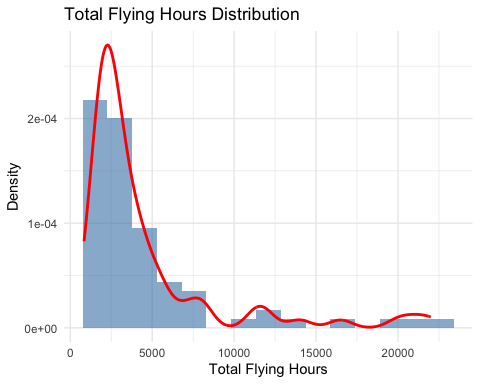
Descriptive Statistics for Total Flying Hours

| n | min\_hours | q1\_hours | median\_hours | mean\_hours | q3\_hours | max\_hours | sd\_hours | iqr\_hours | skew\_hours | kurt\_hours |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 76 | 840 | 2015 | 2900 | 4532.13 | 4500 | 22000 | 4490.88 | 2485 | 2.41 | 5.48 |

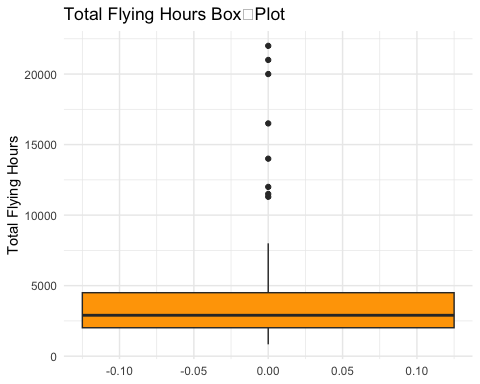
# 3 Shapiro–Wilk normality test ------------------------------------------  
if (hours\_summary$n <= 5000) {  
 shapiro\_p <- shapiro.test(included\_data$total\_flying\_hours)$p.value  
 cat("\nShapiro–Wilk normality test p‑value:",  
 format(shapiro\_p, digits = 4), "\n")  
}

##   
## Shapiro–Wilk normality test p‑value: 3.991e-12

# 4 Histogram with density overlay ---------------------------------------  
ggplot(included\_data, aes(x = total\_flying\_hours)) +  
 geom\_histogram(aes(y = ..density..), bins = 15,  
 fill = "steelblue", alpha = 0.6) +  
 geom\_density(color = "red", linewidth = 1) +  
 labs(title = "Total Flying Hours Distribution",  
 x = "Total Flying Hours", y = "Density") +  
 theme\_minimal()



# 5 Box‑plot for outlier check -------------------------------------------  
ggplot(included\_data, aes(y = total\_flying\_hours)) +  
 geom\_boxplot(fill = "orange", width = 0.25) +  
 labs(title = "Total Flying Hours Box‑Plot",  
 y = "Total Flying Hours") +  
 theme\_minimal()



library(janitor) # tabyl + adorn  
  
# 1 Frequency table with counts & percentages -----------------------------  
gender\_table <- included\_data %>%  
 tabyl(gender) %>% # counts  
 adorn\_totals("row") %>% # add Total row  
 adorn\_pct\_formatting(digits = 1) # add percent column  
  
knitr::kable(gender\_table, caption = "Gender Distribution")

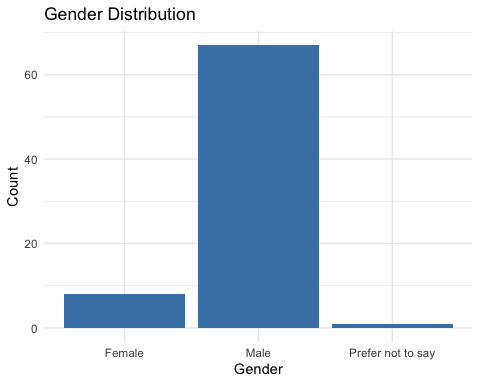
Gender Distribution

| gender | n | percent |
| --- | --- | --- |
| Female | 8 | 10.5% |
| Male | 67 | 88.2% |
| Prefer not to say | 1 | 1.3% |
| Total | 76 | 100.0% |

# 2 Count of missing responses -------------------------------------------  
missing\_gender <- sum(is.na(included\_data$gender) | included\_data$gender == "")  
cat("\nMissing / blank gender responses:", missing\_gender, "\n")

##   
## Missing / blank gender responses: 0

# 3 Bar chart -------------------------------------------------------------  
ggplot(included\_data, aes(x = gender)) +  
 geom\_bar(fill = "steelblue") +  
 labs(title = "Gender Distribution", x = "Gender", y = "Count") +  
 theme\_minimal()



library(janitor) # tabyl + adorn  
  
# 1 Frequency table with counts & percentages -----------------------------  
carrier\_table <- included\_data %>%  
 tabyl(air\_carrier\_type) %>% # counts  
 adorn\_totals("row") %>% # add Total row  
 adorn\_pct\_formatting(digits = 1) # add percent column  
  
knitr::kable(carrier\_table,  
 caption = "Air‑Carrier Type Distribution")

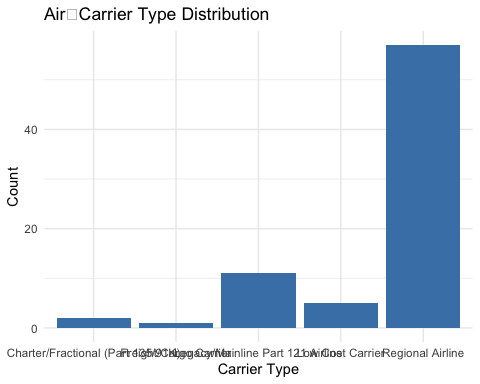
Air‑Carrier Type Distribution

| air\_carrier\_type | n | percent |
| --- | --- | --- |
| Charter/Fractional (Part 135/91K) | 2 | 2.6% |
| Freight/Cargo Carrier | 1 | 1.3% |
| Legacy/Mainline Part 121 Airline | 11 | 14.5% |
| Low Cost Carrier | 5 | 6.6% |
| Regional Airline | 57 | 75.0% |
| Total | 76 | 100.0% |

# 2 Missing responses -----------------------------------------------------  
missing\_carrier <- sum(is.na(included\_data$air\_carrier\_type) |  
 included\_data$air\_carrier\_type == "")  
cat("\nMissing / blank carrier‑type responses:",  
 missing\_carrier, "\n")

##   
## Missing / blank carrier‑type responses: 0

# 3 Bar chart -------------------------------------------------------------  
ggplot(included\_data, aes(x = air\_carrier\_type)) +  
 geom\_bar(fill = "steelblue") +  
 labs(title = "Air‑Carrier Type Distribution",  
 x = "Carrier Type", y = "Count") +  
 theme\_minimal()



library(janitor) # tabyl + adorn  
  
# 1 Frequency table: counts + percentages ------------------------------  
position\_table <- included\_data %>%  
 tabyl(position\_position) %>% # counts  
 adorn\_totals("row") %>% # add Total row  
 adorn\_pct\_formatting(digits = 1) # add % column  
  
knitr::kable(position\_table,  
 caption = "Flight‑Deck Position Distribution")

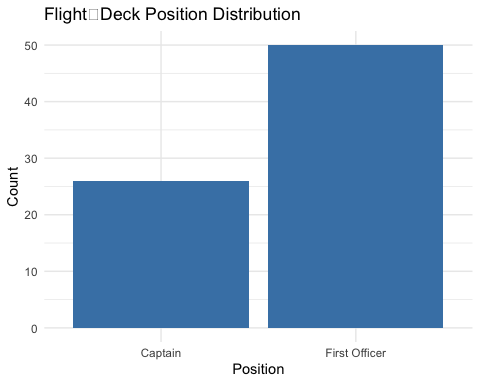
Flight‑Deck Position Distribution

| position\_position | n | percent |
| --- | --- | --- |
| Captain | 26 | 34.2% |
| First Officer | 50 | 65.8% |
| Total | 76 | 100.0% |

# 2 Missing responses ---------------------------------------------------  
missing\_pos <- sum(is.na(included\_data$position\_position) |  
 included\_data$position\_position == "")  
cat("\nMissing / blank position responses:",  
 missing\_pos, "\n")

##   
## Missing / blank position responses: 0

# 3 Bar chart -----------------------------------------------------------  
ggplot(included\_data, aes(x = position\_position)) +  
 geom\_bar(fill = "steelblue") +  
 labs(title = "Flight‑Deck Position Distribution",  
 x = "Position", y = "Count") +  
 theme\_minimal()



library(janitor) # tabyl + adorn  
  
# 1 Recode (two exact strings) -------------------------------------------  
included\_data <- included\_data %>%  
 mutate(  
 military\_bg = case\_when(  
 military\_experience == "Previously served in the military" ~ "Yes",  
 military\_experience == "Never served in the military" ~ "No",  
 TRUE ~ NA\_character\_ # catch unexpected  
 )  
 )  
  
# 2 Frequency table -------------------------------------------------------  
mil\_table <- included\_data %>%  
 tabyl(military\_bg) %>%  
 adorn\_totals("row") %>%  
 adorn\_pct\_formatting(digits = 1)  
  
knitr::kable(mil\_table,  
 caption = "Military Experience Distribution (Yes / No)")

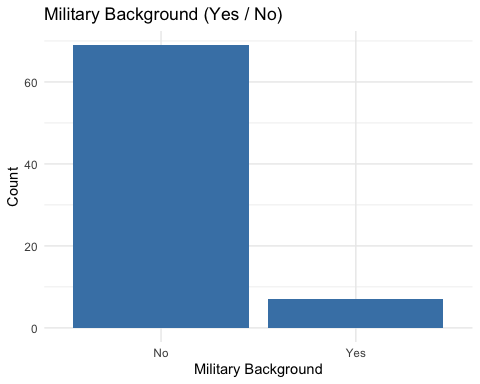
Military Experience Distribution (Yes / No)

| military\_bg | n | percent |
| --- | --- | --- |
| No | 69 | 90.8% |
| Yes | 7 | 9.2% |
| Total | 76 | 100.0% |

# 3 Missing responses -----------------------------------------------------  
missing\_mil <- sum(is.na(included\_data$military\_bg))  
cat("\nMissing military‑experience responses:", missing\_mil, "\n")

##   
## Missing military‑experience responses: 0

# 4 Bar chart -------------------------------------------------------------  
ggplot(included\_data, aes(x = military\_bg)) +  
 geom\_bar(fill = "steelblue") +  
 labs(title = "Military Background (Yes / No)",  
 x = "Military Background", y = "Count") +  
 theme\_minimal()



library(janitor) # tabyl + adorn  
library(stringr) # str\_detect  
  
# ── Part A  Full country table ───────────────────────────────────────────  
country\_table <- included\_data %>%  
 tabyl(nationality) %>% # counts  
 arrange(desc(n)) %>%  
 adorn\_pct\_formatting(digits = 1)  
  
knitr::kable(country\_table,  
 caption = "Nationality – Full Country Breakdown")

Nationality – Full Country Breakdown

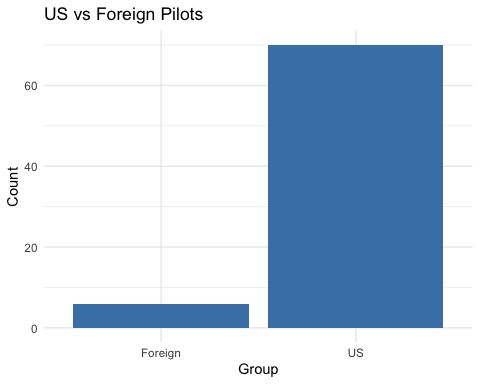
| nationality | n | percent |
| --- | --- | --- |
| United States of America | 70 | 92.1% |
| Ethiopia | 1 | 1.3% |
| Germany | 1 | 1.3% |
| India | 1 | 1.3% |
| Lao People’s Democratic Republic | 1 | 1.3% |
| Mexico | 1 | 1.3% |
| New Zealand | 1 | 1.3% |

# ── Part B  U.S. vs Foreign recode ───────────────────────────────────────  
included\_data <- included\_data %>%  
 mutate(  
 us\_vs\_foreign = case\_when(  
 str\_detect(nationality, regex("^united states", ignore\_case = TRUE)) |  
 str\_detect(nationality, regex("^usa$", ignore\_case = TRUE)) ~ "US",  
 is.na(nationality) | nationality == "" ~ NA\_character\_,  
 TRUE ~ "Foreign"  
 )  
 )  
  
# Counts & percentages  
us\_table <- included\_data %>%  
 tabyl(us\_vs\_foreign) %>%  
 adorn\_totals("row") %>%  
 adorn\_pct\_formatting(digits = 1)  
  
knitr::kable(us\_table,  
 caption = "Nationality Re‑coded to US vs Foreign")

Nationality Re‑coded to US vs Foreign

| us\_vs\_foreign | n | percent |
| --- | --- | --- |
| Foreign | 6 | 7.9% |
| US | 70 | 92.1% |
| Total | 76 | 100.0% |

# Bar chart  
ggplot(included\_data, aes(x = us\_vs\_foreign)) +  
 geom\_bar(fill = "steelblue") +  
 labs(title = "US vs Foreign Pilots",  
 x = "Group", y = "Count") +  
 theme\_minimal()



## NEXT STEPS

Descriptive summaries (age, gender, etc.) Mann–Whitney U and Kruskal–Wallis tests Visualisations (box‑plots, violin plots, etc.)