**R Scripts**

* **Summarized stats for Zillow and Realtor**

# **Load required libraries**

library(dplyr) library(lubridate) library(writexl)

# **Step 1: Extract year from the Date column**

housing\_data <- housing\_data %>% mutate(Year = year(as.Date(Date)))

# **Step 2: Filter for years between 2016 and 2025**

housing\_filtered <- housing\_data %>% filter(Year >= 2016 & Year <= 2025)

# **Step 3: Define columns to summarize**

cols\_to\_summarize <- c( "Avg Listing Price - Realtor", "Avg Active Inventory - Realtor", "Avg Days on Market- Realtor", "Avg Pending Inventory - Realtor", "Avg Listing Price - Zillow", "Avg Active Inventory - Zillow", "Avg Days on Market - Zillow", "Avg Pending Inventory - Zillow" )

# **Step 4: Define safe summary functions**

safe\_min <- function(x) if (all(is.na(x))) NA\_real\_ else min(x, na.rm = TRUE) safe\_max <- function(x) if (all(is.na(x))) NA\_real\_ else max(x, na.rm = TRUE) safe\_median <- function(x) if (all(is.na(x))) NA\_real\_ else median(x, na.rm = TRUE) safe\_sd <- function(x) if (sum(!is.na(x)) > 1) sd(x, na.rm = TRUE) else NA\_real\_

# **Step 5: Summarize data by State and Year (no mean, no count)**

summary\_stats\_by\_year <- housing\_filtered %>% group\_by(State Abbr, Year) %>% summarize(across(all\_of(cols\_to\_summarize), list( median = safe\_median, sd = safe\_sd, min = safe\_min, max = safe\_max ), .names = "{.col}\_{.fn}"), .groups = "drop") %>% filter(if\_any(everything(), ~ !is.na(.))) # remove rows where all summary stats are NA

# **Step 6: Export to Excel**

write\_xlsx(summary\_stats\_by\_year, "housing\_summary\_by\_year\_2016\_2025.xlsx")

Changed name to Housing KPI Summary (2016-2025) after further refinement and validation using Excel

* **Calculated correlations for Zillow & Realtor**

library(dplyr) library(lubridate) library(tidyr)

# **Step 1: Extract Year from Date**

housing\_data <- housing\_data %>% mutate(Date = as.Date(Date), Year = year(Date))

# **Step 2: Filter for 2016 to 2025**

housing\_filtered <- housing\_data %>% filter(Year >= 2016 & Year <= 2025)

# **Step 3: Clean numeric columns (remove $ and commas, convert to numeric)**

housing\_clean <- housing\_filtered %>% mutate(across( c( "Avg Listing Price - Realtor", "Avg Active Inventory - Realtor", "Avg Days on Market- Realtor", "Avg Pending Inventory - Realtor", "Avg Listing Price - Zillow", "Avg Active Inventory - Zillow", "Avg Days on Market - Zillow", "Avg Pending Inventory - Zillow" ), ~ as.numeric(gsub("[\$,]", "", .)) ))

# **Step 4: Function to compute correlation excluding Zillow vs Realtor**

compute\_cor\_long <- function(df) { num\_df <- df %>% select( "Avg Listing Price - Realtor", "Avg Active Inventory - Realtor", "Avg Days on Market- Realtor", "Avg Pending Inventory - Realtor", "Avg Listing Price - Zillow", "Avg Active Inventory - Zillow", "Avg Days on Market - Zillow", "Avg Pending Inventory - Zillow" )

# **Remove columns with NA or 0 standard deviation**

valid\_cols <- sapply(num\_df, function(x) { s <- sd(x, na.rm = TRUE) !is.na(s) && s > 0 }) num\_df <- num\_df[, valid\_cols, drop = FALSE]

if (ncol(num\_df) < 2) return(tibble(Var1 = character(), Var2 = character(), Correlation = numeric()))

cor\_mat <- cor(num\_df, use = "pairwise.complete.obs")

cor\_df <- as.data.frame(as.table(cor\_mat)) %>% rename(Var1 = Var1, Var2 = Var2, Correlation = Freq) %>% filter(Var1 != Var2) %>% distinct()

# **Exclude Zillow vs Realtor combinations**

cor\_df\_filtered <- cor\_df %>% filter( !(grepl("Zillow", Var1) & grepl("Realtor", Var2)) & !(grepl("Realtor", Var1) & grepl("Zillow", Var2)) )

return(cor\_df\_filtered) }

# **Step 5: Group by State and Year and calculate correlations**

correlation\_results <- housing\_clean %>% group\_by(State Abbr, Year) %>% group\_modify(~ compute\_cor\_long(.x)) %>% ungroup()

# **Step 6: View or export**

print(head(correlation\_results))