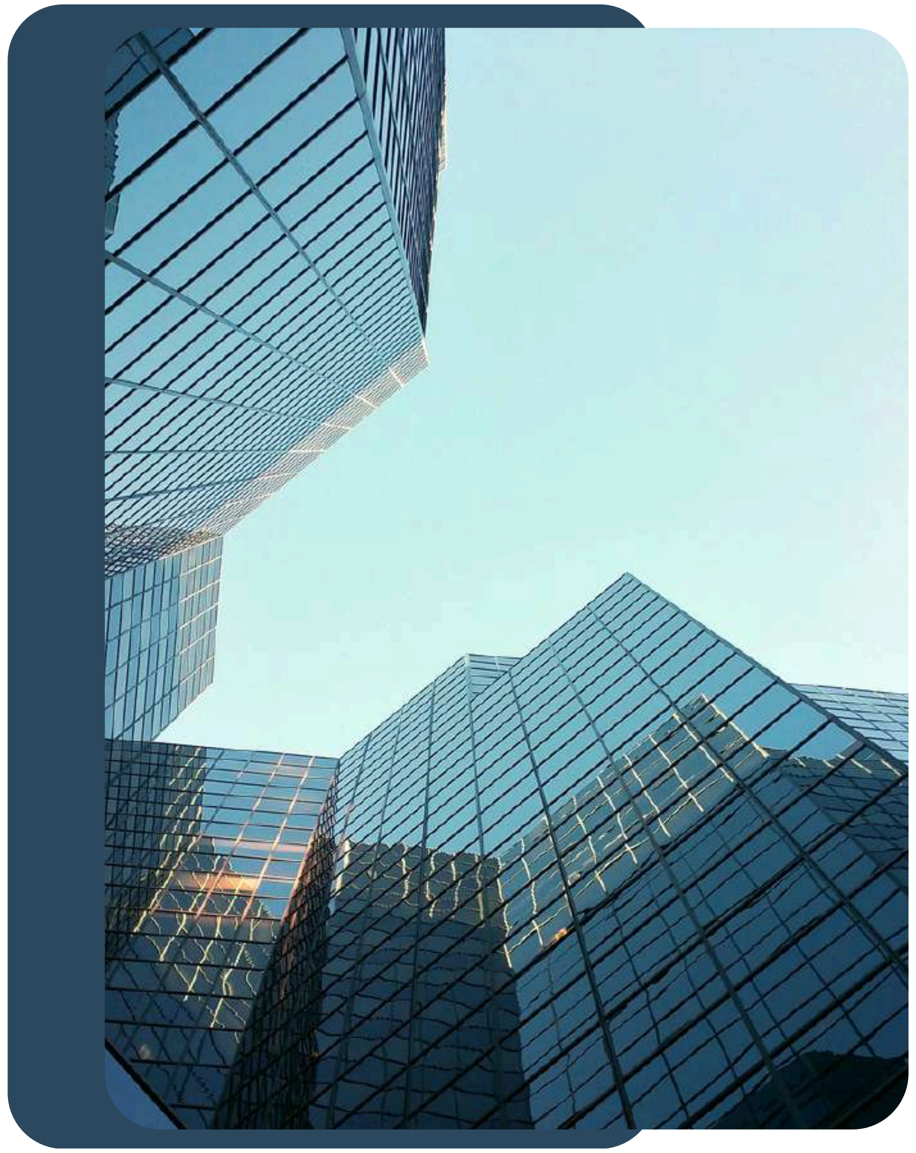
Friday, 9th May 2025

NICHOLAS FEDLIM
3036385482

What Defines Hong Kong's Job Market Health?

HKU - COMP2501

Table Of Content



PART 1:

INTRODUCTION

Introduction

Questions to Answer

1. How have key job market indicators (e.g. **unemployment rate, median salary, job vacancies**) trended from 2019 to 2024, and which sectors (e.g., finance, retail, tech) show the most significant changes?
2. Which job market indicators (**unemployment rate, median salary, or vacancies**) most reliably reflect HK's economic health, and how does their predictive value (Trends) vary across years?

The Importance

- **Career Insights for Students:** Analyses 2019–2024 job market trends (e.g., salaries, unemployment) to help undergrads identify opportunities in sectors like finance or tech.
- **Guiding Sectoral Strategies:** Identifies significant changes in sectors like finance and retail, which will support targeted job creation and investment decisions.

RAW DATASETS USED

Main Collection Source: https://data.gov.hk/en/

Unemployment Rate

- Table 210-06103 : Unemployment rate and underemployment rate by age and sex
- Table 210-06101 : Statistics on labour force, employment, unemployment and underemployment

Median Salary

- Table 210-06316 : Median monthly employment earnings of employed persons by industry of main employment and sex
- Nominal Wages Trends in Hong Kong.csv
- Real Wages Trends in Hong Kong.csv

Job Vacancies

- Table 215-16007 : Number of vacancies (other than those in the civil service) analysed by industry section and major occupation group

Cleaning & Preparing the Data, How?

1. Read the CSV File

```r
# Prompt user to select the CSV file
cat("Please select the 'Jobs Vacancies HK.csv' file\n")
file_path <- file.choose()

# Reading the CSV file
vacancies_data <- read_csv(file_path,
                           skip = 5)  # Skipping title, empty rows, and header rows
```

2. Remove Useless rows and renaming columns

```r
# Remove empty rows and notes at the bottom
vacancies_data <- vacancies_data %>%
  filter(!is.na(`...1`), filter(!str_detect(`...1`, "Note|Source|Release Date")))

# Rename columns properly based on the header structure
colnames(vacancies_data) <- c("Year", "Month", "Industry_Section", "Managers",
                              "Professionals", "Associate_Professionals",
                              "Clerical_Support_Workers", "Service_Sales_Workers",
                              "Craft_Related_Workers", "Plant_Machine_Operators",
                              "Skilled_Agricultural_Fishery", "Elementary_Occupations", "Total")
```

3. Cleaning the Data (Remove Chinese Characters)

```r
# Get the English column headers (years and rate)
headers <- raw_data[1, ] %>%
  as.character() %>%
  str_replace_all("\n.*", "") %>%  # Remove Chinese parts
  str_trim() %>%
  make_clean_names()
```

4. Cleaning the Data (Mutate, Filter, Select, etc.)

```r
# Clean the data
vacancies_data_clean <- vacancies_data %>%
 # Convert Year to numeric
 mutate(Year = as.numeric(Year)) %>%
 # Clean Month names
 mutate(Month = case_when(Month == "Mar" ~ "March", Month == "Jun" ~ "June",
                          Month == "Sep" ~ "September", Month == "Dec" ~ "December",
                          TRUE ~ Month)) %>%
 # Clean Industry_Section names
 mutate(Industry_Section = case_when(
     Industry_Section == "C: Manufacturing" ~ "Manufacturing",
     Industry_Section == "B, D & E: Mining and quarrying; and electricity and gas supply, and waste management"
                        ~ "Mining, Quarrying, Electricity, Gas, Waste Management",
     Industry_Section == "F: Construction sites (manual workers only)" ~ "Construction Sites (Manual Workers)",
     # Add all other industry sections similarly
     TRUE ~ Industry_Section ) ) %>%
 # Convert all numeric columns to numeric (handling "-" as NA)
 mutate(across(4:13, ~ifelse(. == "-", NA, as.numeric(.))))
```

5. Repeat the four steps for 5 other data

Cleaning & Preparing the Data

From: RAW DATA

	X1	X2	X3	X4	X5	X6
1	NA	2014	2015	NA	2016	NA
2	行業主類 Industry section	指數 Index	指數 Index	按年變動率 Year-on-year rate of change (%)	指數 Index	按年變動率 Year-on-year rate
3	製造 Manufacturing	189.7	196.6	3.6	203.6	3.6
4	進出口貿易、批發及零售 Import/export, wholesale and …	203.1	209.7	3.3	215.5	2.8
5	運輸 Transportation	177.6	185.8	4.6	192.5	3.6
6	住宿及膳食服務活動 Accommodation and food service …	174.4	184	5.5	193.1	4.9
7	金融及保險活動 Financial and insurance activities	212	218.4	3	225.5	3.2
8	地產租賃及保養管理 Real estate leasing and maintenan…	222.1	230.7	3.9	238.6	3.4
9	專業及商業服務 Professional and business services	215.6	231.7	7.4	242.6	4.7
10	個人服務 Personal services	261.1	277.6	6.3	293.7	5.8
11	所有選定行業主類@ All selected industry sections@	200.3	209.4	4.6	217.3	3.8
12	NA	NA	NA	NA	NA	NA
13	註釋:	指有關年度6月的工資指數。	NA	NA	NA	NA
14	NA	按年變動率以小數點後兩個位的工資指數計算。	NA	NA	NA	NA
15	NA	@ 指「勞工收入統計調查」中工資調查涵蓋的所有行業，包…	NA	NA	NA	NA
16	Notes:	Figures refer to wage indices for June of the year.	NA	NA	NA	NA
17	NA	Year-on-year rates of change are derived from wage i…	NA	NA	NA	NA
18	NA	@ Refers to all industries covered by the wage enquir…	NA	NA	NA	NA

```r
# Loading required libraries
library(tidyverse)
library(dplyr)
library(tidyr)
library(stringr)
library(readr)
```

To: Cleaned Data

	Industry	2014	2015	2015 YoY Rate of Change (%)	2016	2016 YoY Rate of Change (%)	2017	2017 YoY Rate of Change (%)	2018
1	Manufacturing	189.7	196.6	3.6	203.6	3.6	211.5	3.9	21
2	Import/export, wholesale and retail trades	203.1	209.7	3.3	215.5	2.8	221.7	2.9	22
3	Transportation	177.6	185.8	4.6	192.5	3.6	198.3	3.0	20
4	Accommodation and food service activities	174.4	184.0	5.5	193.1	4.9	202.5	4.9	21
5	Financial and insurance activities	212.0	218.4	3.0	225.5	3.2	233.7	3.6	24
6	Real estate leasing and maintenance management	222.1	230.7	3.9	238.6	3.4	248.2	4.0	25
7	Professional and business services	215.6	231.7	7.4	242.6	4.7	255.0	5.1	26
8	Personal services	261.1	277.6	6.3	293.7	5.8	307.2	4.6	31

PART 2:

METHODOLOGY & FINDINGS

Exploratory Data Analysis (EDA), How?

1. Load the necessary Library, set working directory, and load all of the cleaned data

```r
# Loading required libraries
library(tidyverse)   # For data manipulation and visualization
library(ggplot2)     # For plotting
library(corrplot)    # For correlation analysis
library(forecast)    # For time series analysis
library(lubridate)   # For date handling
library(gridExtra)   # For arranging multiple plots
library(readr)       # For reading CSV files
```

2. Providing all the necessary summary statistics

```r
# Function to summarize dataset
summarize_dataset <- function(data, name) {
  cat("\n--- Summary of", name, "---\n")
  print("Structure:")
  str(data)
  print("Summary Statistics:")
  if (ncol(data) > 0) {
    summary_data <- data %>% summarise(across(where(is.numeric),
                                          list(mean = mean, median = median, sd = sd), na.rm = TRUE))
    print(summary_data)
    print(summary(data))
  } else {
    cat("No columns to summarize.\n")
  }
  print("Missing Values:")
  print(colSums(is.na(data)))
  print("Duplicate Rows:")
  print(sum(duplicated(data)))
}
```

3. Finding the Outliers using interquartile (IQR) method

```r
# Function to check outliers (using IQR method)
check_outliers <- function(data, column, name) {
  q <- quantile(data[[column]], c(0.25, 0.75), na.rm = TRUE)
  iqr <- q[2] - q[1]
  lower_bound <- q[1] - 1.5 * iqr
  upper_bound <- q[2] + 1.5 * iqr
  outliers <- sum(data[[column]] < lower_bound | data[[column]] > upper_bound, na.rm = TRUE)
  cat("Outliers in", column, "of", name, ":", outliers, "\n")
}
```

4. Applying all the function

```r
# Applying summary function to all datasets
summarize_dataset(earnings_data, "Earnings Data")
summarize_dataset(vacancies_data, "Vacancies Data")
summarize_dataset(unemployment_data, "Unemployment Data")
summarize_dataset(labor_data, "Labor Data")
summarize_dataset(real_wages_data, "Real Wages Data")
summarize_dataset(nominal_wages_data, "Nominal Wages Data")

# Checking outliers for key numeric columns
check_outliers(earnings_data, "both_earnings", "Earnings Data")
check_outliers(unemployment_data, "unemployment_rate", "Unemployment Data")
check_outliers(real_wages_data, "wage_index", "Real Wages Data")
check_outliers(nominal_wages_data, "wage_index", "Nominal Wages Data")
```

5. Use ggplot2 to make the relevant data visualization

Exploratory Data Analysis (EDA)

1. Outliers Result

```
> # Checking outliers for key numeric columns
> check_outliers(earnings_data, "both_earnings", "Earnings Data")
Outliers in both_earnings of Earnings Data : 0
> check_outliers(unemployment_data, "unemployment_rate", "Unemployment Data")
Outliers in unemployment_rate of Unemployment Data : 0
> check_outliers(real_wages_data, "wage_index", "Real Wages Data")
Outliers in wage_index of Real Wages Data : 0
> check_outliers(nominal_wages_data, "wage_index", "Nominal Wages Data")
Outliers in wage_index of Nominal Wages Data : 0
```

What is Exploratory Data Analysis?

- Exploratory data analysis (EDA) is used **to analyze and investigate** data sets and **summarize** their main characteristics.
- The main purpose of EDA is to help look at data before making any assumptions.

2. Summary Statistics Result (One of the datasets)

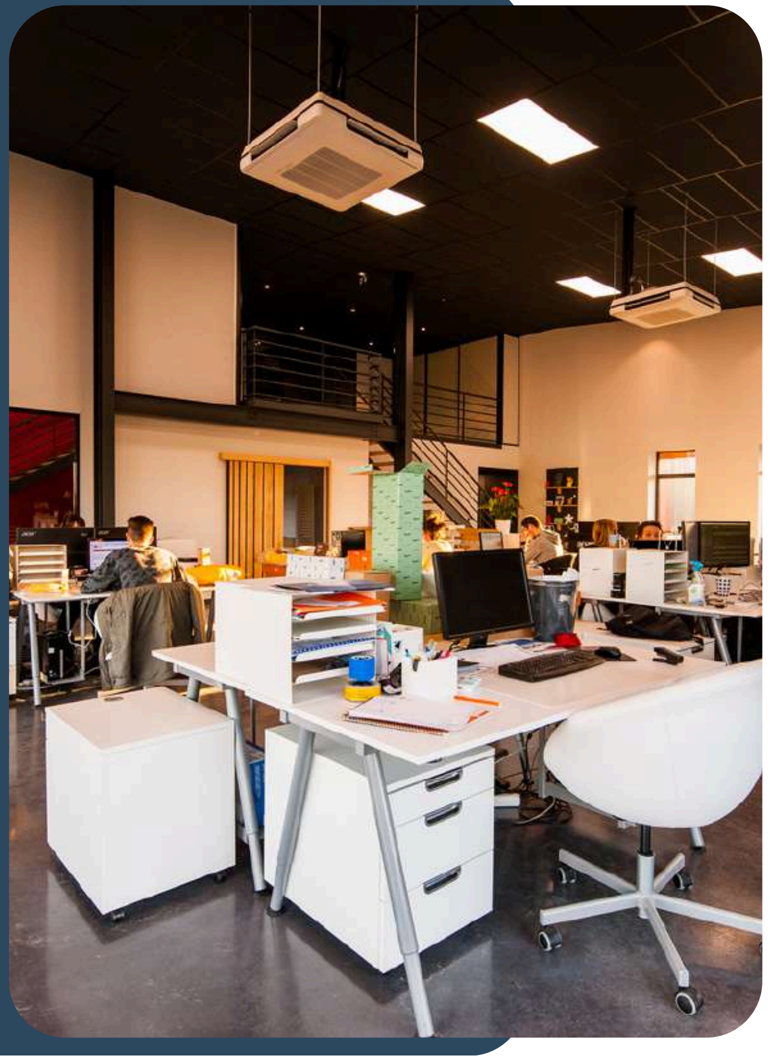
```
[1] "Summary Statistics:"
# A tibble: 1 × 66
  `2014_mean` `2014_median` `2014_sd` `2015_mean` `2015_median` `2015_sd` 2015 YoY Rate of Change (…¹ 2015 YoY Rate of Cha…²
        <dbl>         <dbl>     <dbl>       <dbl>         <dbl>     <dbl>                       <dbl>                  <dbl>
1        124.          124.      16.7        124.          123.      17.6                       0.413                  -0.05
# ℹ abbreviated names: ¹`2015 YoY Rate of Change (%)_mean`, ²`2015 YoY Rate of Change (%)_median`
# ℹ 58 more variables: `2015 YoY Rate of Change (%)_sd` <dbl>, `2016_mean` <dbl>, `2016_median` <dbl>, `2016_sd` <dbl>,
#   `2016 YoY Rate of Change (%)_mean` <dbl>, `2016 YoY Rate of Change (%)_median` <dbl>,
#   `2016 YoY Rate of Change (%)_sd` <dbl>, `2017_mean` <dbl>, `2017_median` <dbl>, `2017_sd` <dbl>,
#   `2017 YoY Rate of Change (%)_mean` <dbl>, `2017 YoY Rate of Change (%)_median` <dbl>,
#   `2017 YoY Rate of Change (%)_sd` <dbl>, `2018_mean` <dbl>, `2018_median` <dbl>, `2018_sd` <dbl>,
#   `2018 YoY Rate of Change (%)_mean` <dbl>, `2018 YoY Rate of Change (%)_median` <dbl>, …
   Industry              2014            2015        2015 YoY Rate of Change (%)      2016        2016 YoY Rate of Change (%)
 Length:8           Min.   :104.3   Min.   :105.5   Min.   :-1.2000              Min.   :107.7   Min.   :0.300
 Class :character   1st Qu.:111.6   1st Qu.:111.2   1st Qu.:-0.7000              1st Qu.:112.4   1st Qu.:0.850
 Mode  :character   Median :124.1   Median :122.7   Median :-0.0500              Median :123.3   Median :1.050
                    Mean   :123.7   Mean   :124.3   Mean   : 0.4125              Mean   :126.2   Mean   :1.462
                    3rd Qu.:129.9   3rd Qu.:132.4   3rd Qu.: 1.4000              3rd Qu.:134.1   3rd Qu.:2.175
                    Max.   :156.1   Max.   :159.2   Max.   : 3.0000              Max.   :164.3   Max.   :3.200

      2017       2017 YoY Rate of Change (%)      2018       2018 YoY Rate of Change (%)      2019
 Min.   :108.4   Min.   :0.500               Min.   :109.0   Min.   :0.400               Min.   :110.6
 1st Qu.:114.4   1st Qu.:1.150               1st Qu.:116.0   1st Qu.:0.900               1st Qu.:116.0
 Median :124.5   Median :1.600               Median :125.3   Median :1.300               Median :124.5
 Mean   :128.3   Mean   :1.637               Mean   :130.0   Mean   :1.288               Mean   :130.1
 3rd Qu.:136.6   3rd Qu.:2.275               3rd Qu.:139.1   3rd Qu.:1.800               3rd Qu.:139.3
 Max.   :167.9   Max.   :2.700               Max.   :170.3   Max.   :2.100               Max.   :170.4

 2019 YoY Rate of Change (%)      2020       2020 YoY Rate of Change (%)      2021       2021 YoY Rate of Change (%)
 Min.   :-1.3000              Min.   :110.5   Min.   :-0.2000              Min.   :108.1   Min.   :-2.2000
 1st Qu.:-0.1000              1st Qu.:117.0   1st Qu.: 0.0500              1st Qu.:117.0   1st Qu.:-0.9500
 Median : 0.1000              Median :125.8   Median : 0.3500              Median :126.0   Median :-0.5000
 Mean   : 0.1143              Mean   :130.9   Mean   : 0.6625              Mean   :130.7   Mean   :-0.2857
 3rd Qu.: 0.4000              3rd Qu.:141.4   3rd Qu.: 1.3250              3rd Qu.:142.8   3rd Qu.: 0.6000
 Max.   : 1.4000              Max.   :170.1   Max.   : 1.9000              Max.   :168.0   Max.   : 1.4000
```

Result Findings 1

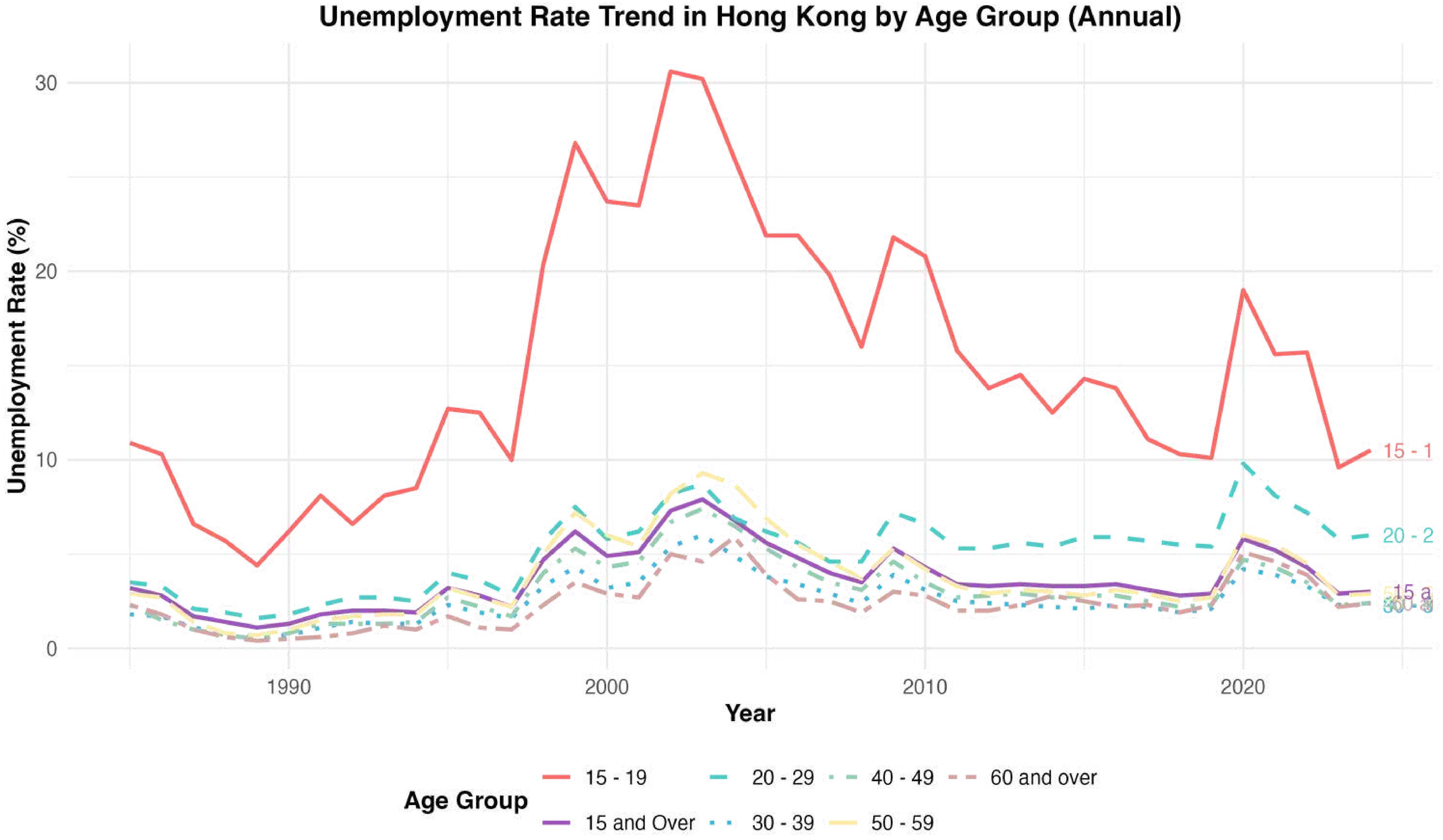
--- Trends (2019–2024) ---



- **Unemployment Rate:** Spiked in 2020–2021 (e.g., 5–7%) due to COVID-19, then declined by 2024, with retail and accommodation sectors hit hardest.
- **Median Salary:** Grew steadily, with Financing and Insurance reaching 38,100 HK$ by 2024, while Retail lagged at 15,200 HK$, reflecting sector-specific recoveries.
- **Job Vacancies:** Dropped in 2020–2021 but rose by 2024, with Financing and Insurance leading (~5,000 vacancies) and Retail trailing (~1,500–2,000).

Data Visualization 1

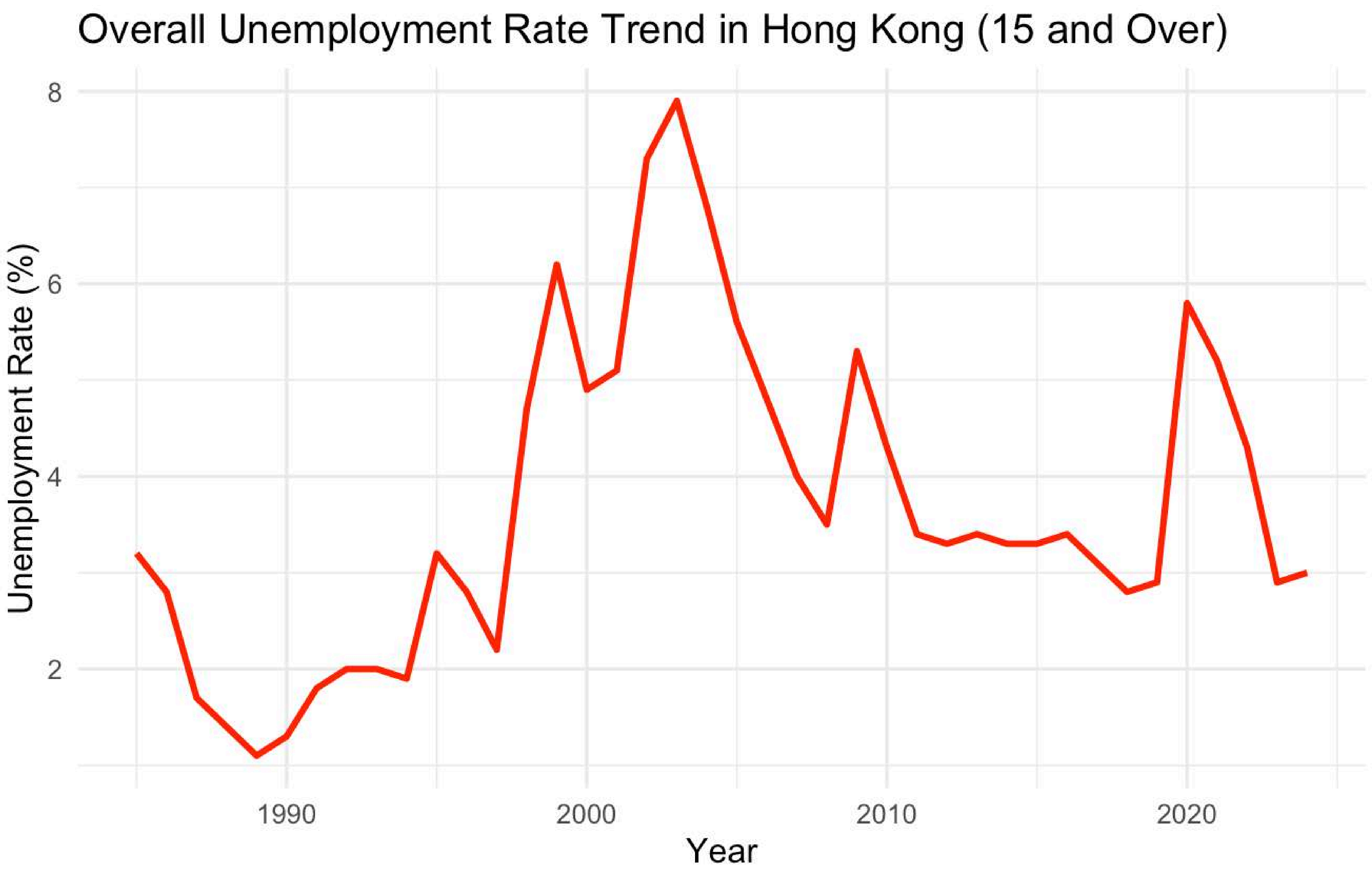
1. Unemployment Rate Trend by Age Group (Line Plot)



First, I filter the data for annual trends and both sexes unemployment rate only. Then, I group by age groups, and use distinct colours and line types to better differentiate each age group.

Data Visualization 1a

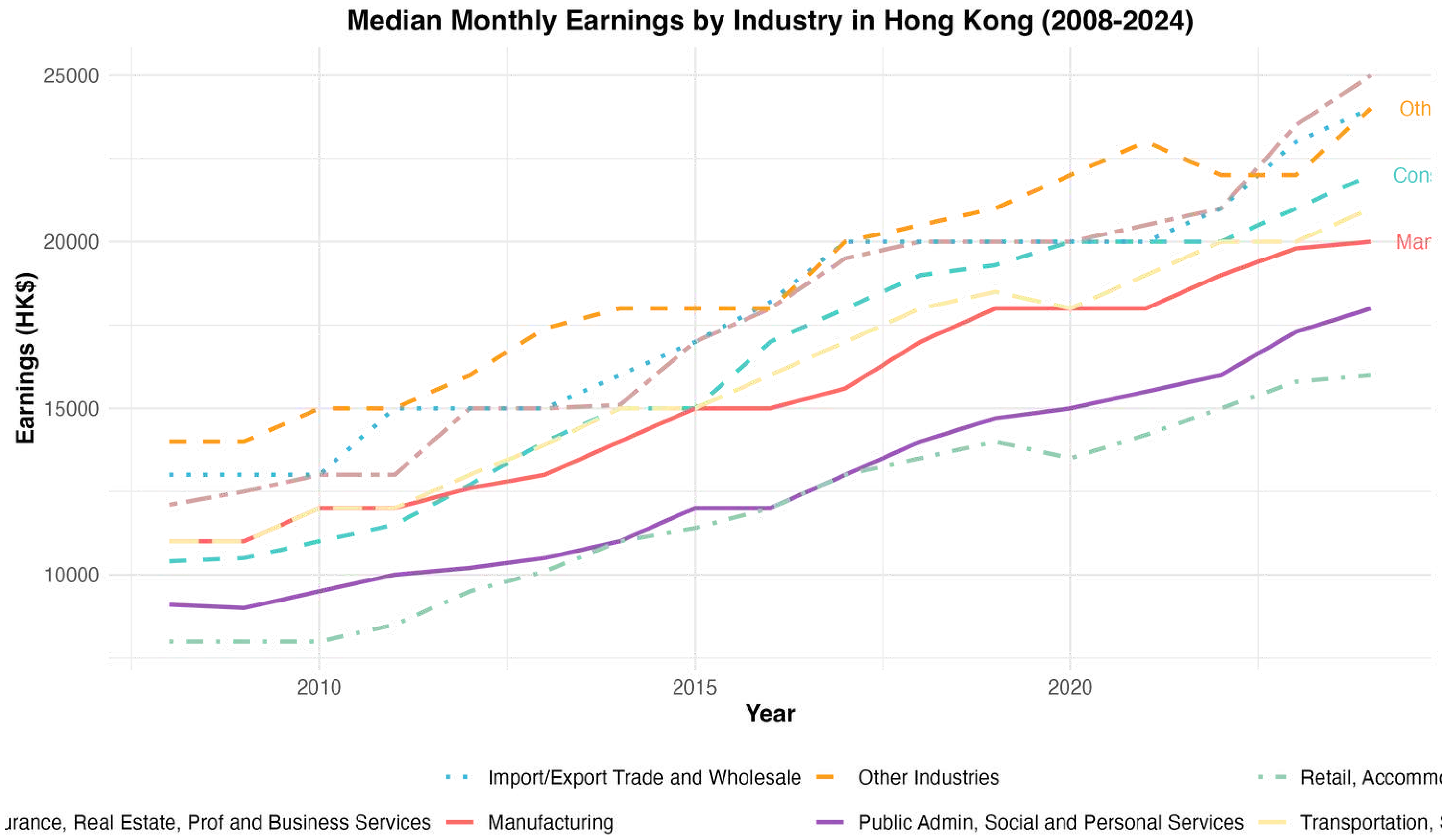
1a. **Overall** Unemployment Rate Trend (Line Plot)



First, I filter the data for annual trends and both sexes unemployment rate only. Then, I filter to include the age group of 15 and above to see the overall trend.

Data Visualization 2

2. Earnings by Industry Over Time (Line Plot)



First, I define the main industry to be analysed. I exclude several industries that overlap with each other. Then, I make the line plot. I use distinct color and line types to better differentiate each industry.

Data Visualization 3

3. Real vs Nominal Wages (Line Plot)

First, load and reshape the nominal and real wages data using pivot_longer(), then I combine the two datasets into one, using bind_rows(). I create the line plot, differentiating between financial activities (Continuous line) and retail activities (Dotted line). Considering that Hong Kong is best known for its financial market.

Nominal VS Real Wages

Nominal

The actual amount of money a worker earns, expressed in current dollars (or any currency) **without adjusting for inflation.**

It reflects the **face value of earnings**, not actual purchasing power.

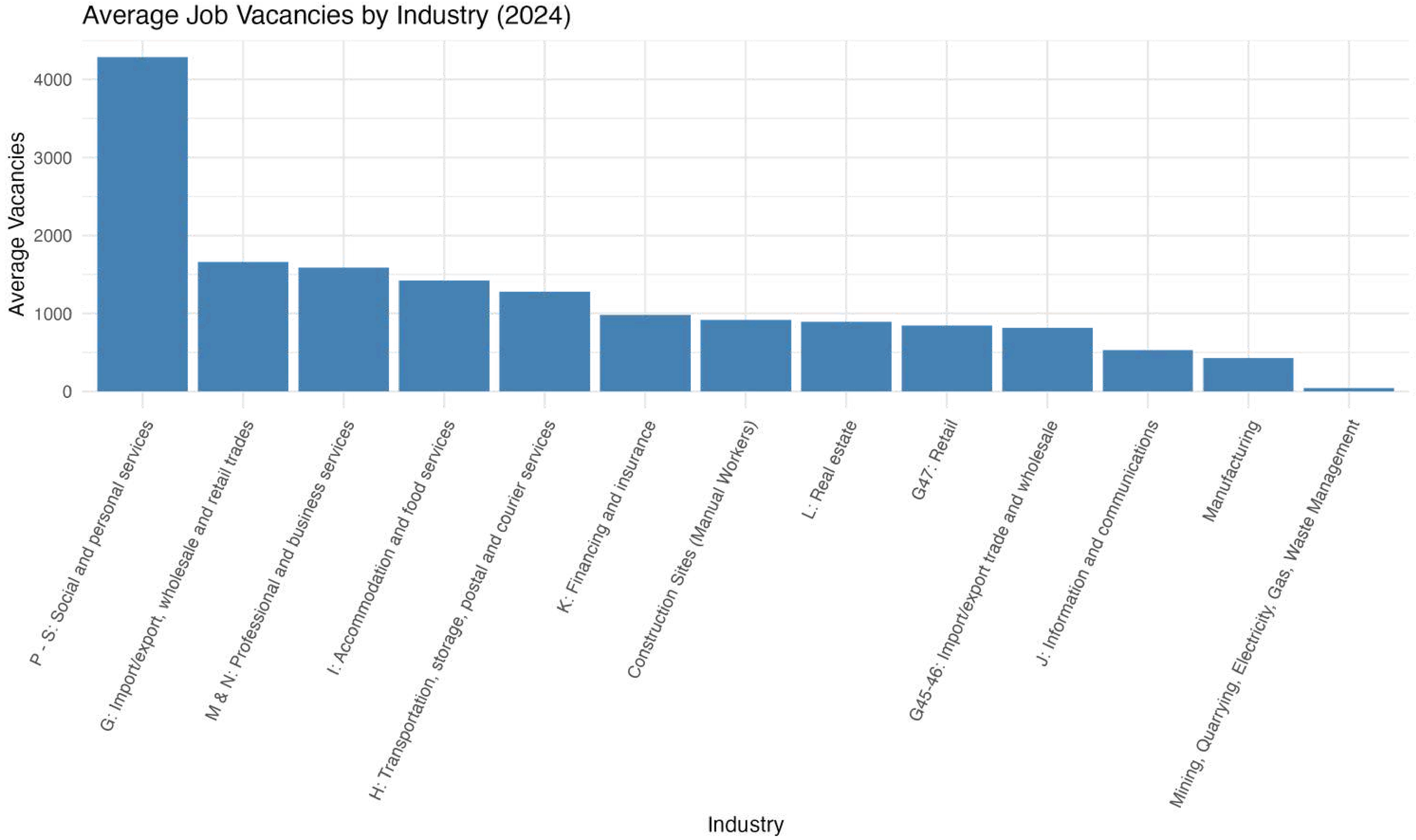
Real

The purchasing power of wages **after adjusting for inflation (or deflation)**.

It measures how many goods and services the wage can actually buy (**Actual standard of living**).

Data Visualization 4

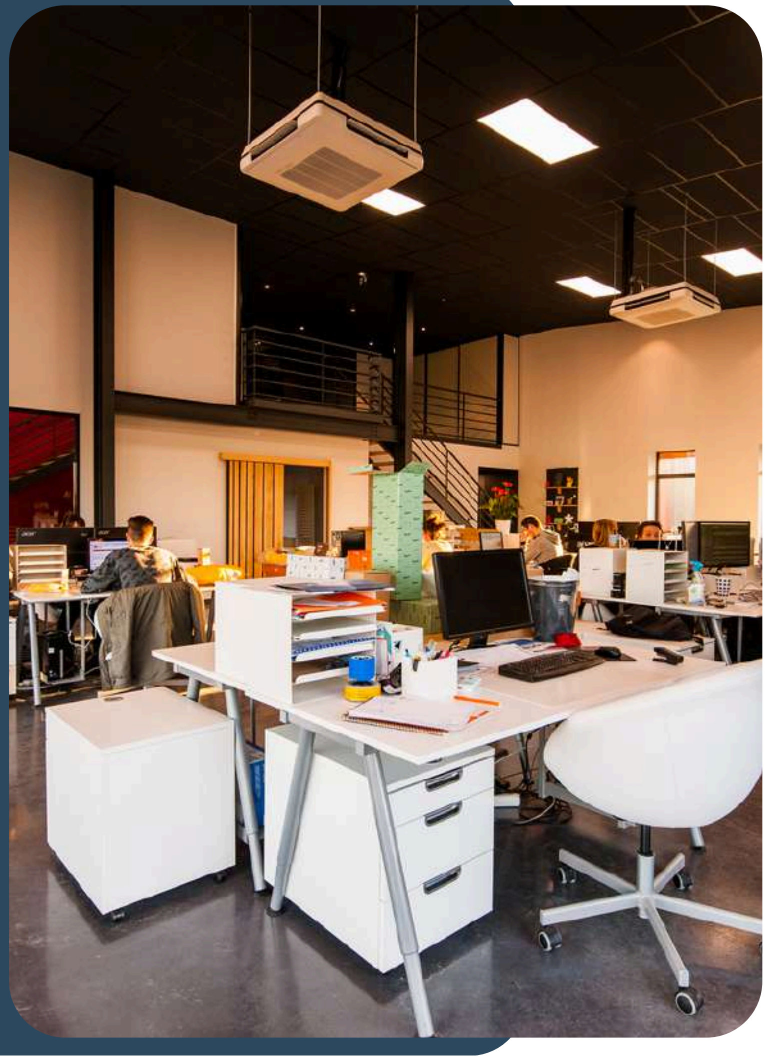
4. Vacancies by Industry (Bar Plot for Latest Year - 2024)



First, I filter the data to only include the latest year in my observation (2024), I reshape the data using pivot_longer(), group by industry section, and find the average vacancies.

Result Findings 2

--- Significant Sectoral Changes ---



- **Finance/Insurance:** Most significant increase in earnings (38,100 HK$), vacancies (~5,000), and wage growth (0.99 correlation with nominal wage), driven by financial hub status.
- **Retail:** Most significant decline or stagnation in earnings (15,200 HK$), vacancies (~1,500–2,000), and wage growth, due to e-commerce and tourism drops.
- **Tech (inferred):** Moderate growth in vacancies and earnings (e.g., ~2,000–3,000 vacancies, 20,000–25,000 HK$ earnings), suggesting emerging demand.

Sectoral Salary Analysis, How?

How to Make the Sectoral Analysis Data:

1. Filter for the latest year, exclude "Overall", and compute the gender pay gap.
2. Check for NA values in sectoral_analysis

How to Make the Sectoral Analysis Plot Data:

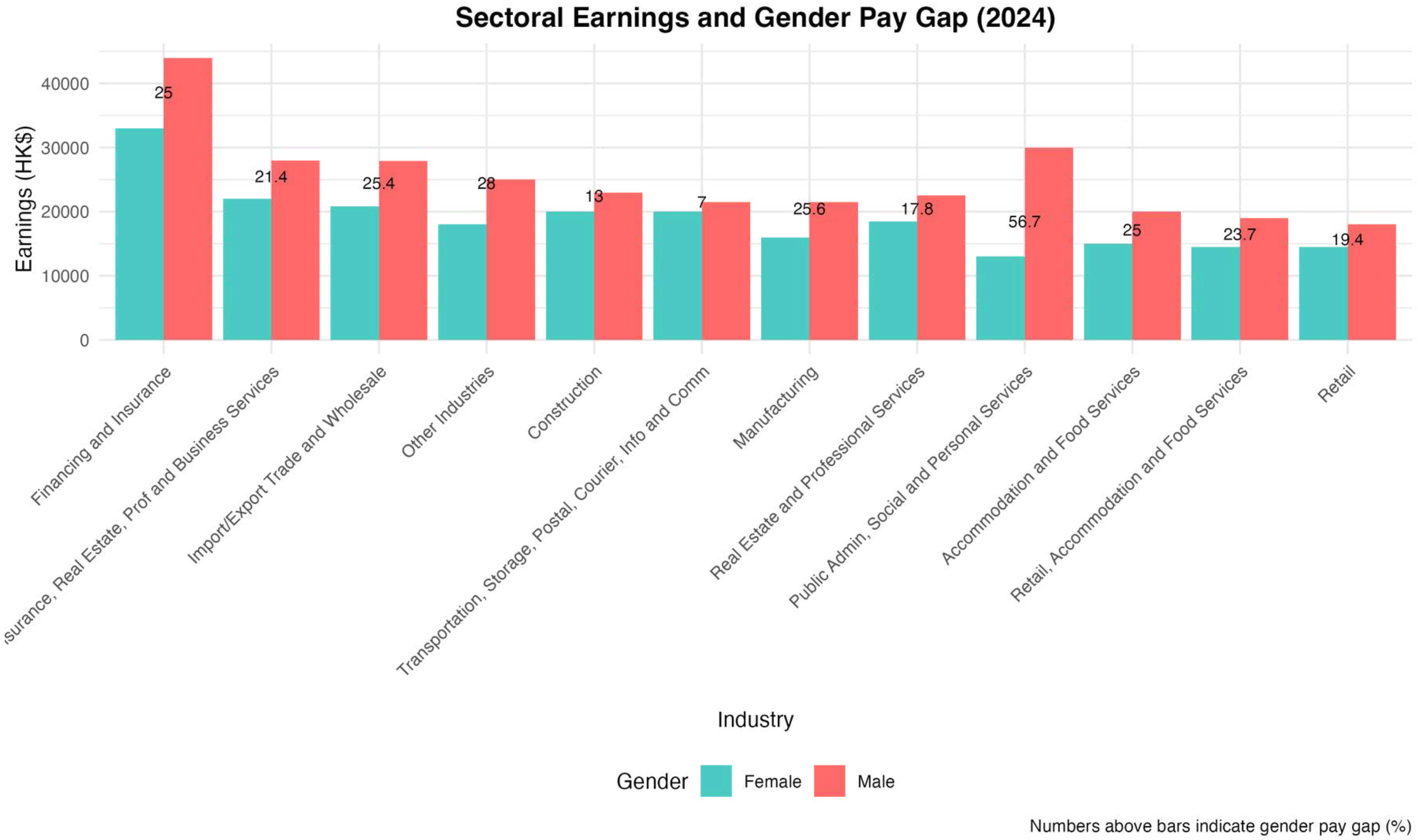
1. Reshape the sectoral analysis data for plotting male and female earnings.
2. Check the reshaped data
3. Visualize sectoral earnings with gender comparison

sectoral_salary_analysis

Industry	avg_earnings	male_earnings	female_earnings	gender_pay_gap
Financing and Insurance	38100	44000	33000	25
Finance, Insurance, Real Estate, Prof and Business Services	25000	28000	22000	21.428571428571400
Import/Export Trade and Wholesale	24000	27900	20800	25.448028673835100
Other Industries	24000	25000	18000	28.000000000000000
Construction	22000	23000	20000	13.043478260869600
Transportation, Storage, Postal, Courier, Info and Comm	21000	21500	20000	6.976744186046510
Manufacturing	20000	21500	16000	25.581395348837200
Real Estate and Professional Services	20000	22500	18500	17.77777777777780
Public Admin, Social and Personal Services	18000	30000	13000	56.66666666666700
Accommodation and Food Services	16700	20000	15000	25
Retail, Accommodation and Food Services	16000	19000	14500	23.684210526315800
Retail	15200	18000	14500	19.444444444444400

Final Result of the Data

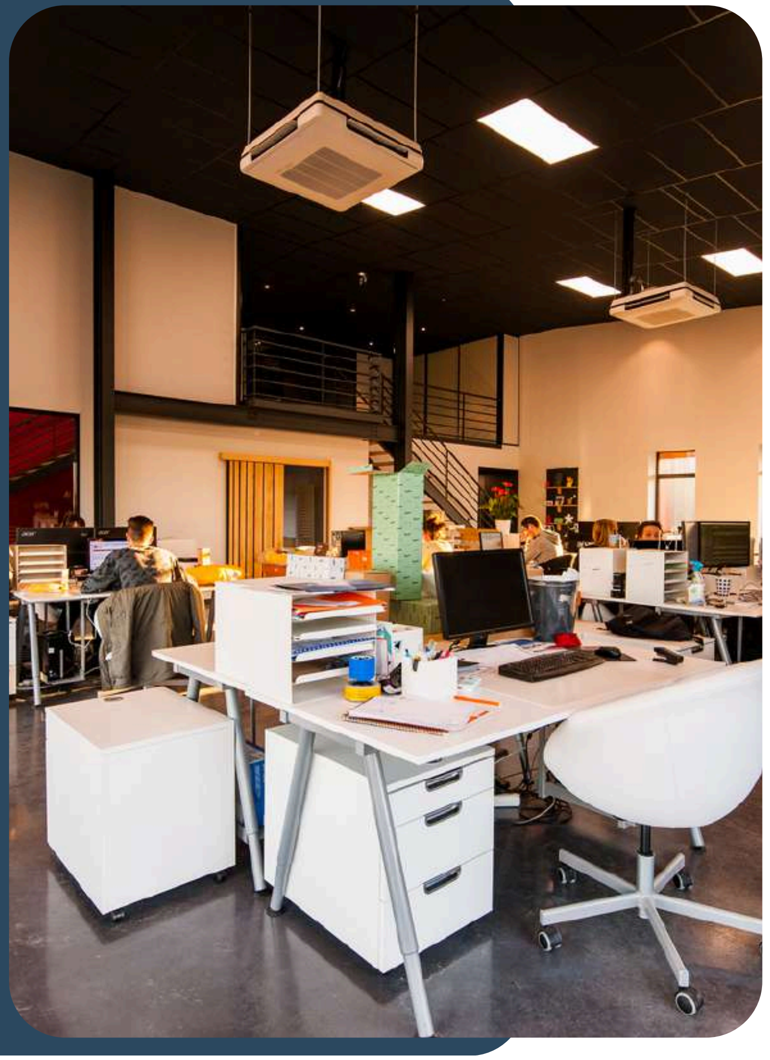
Sectoral Salary Analysis



First, I find the average earnings, male earnings, female earnings, and the gender pay gap. Then, I reshape the data using pivot_longer(). Lastly, I visualize it using geom_bar().

Result Findings 3

--- Reliability of Indicators ---



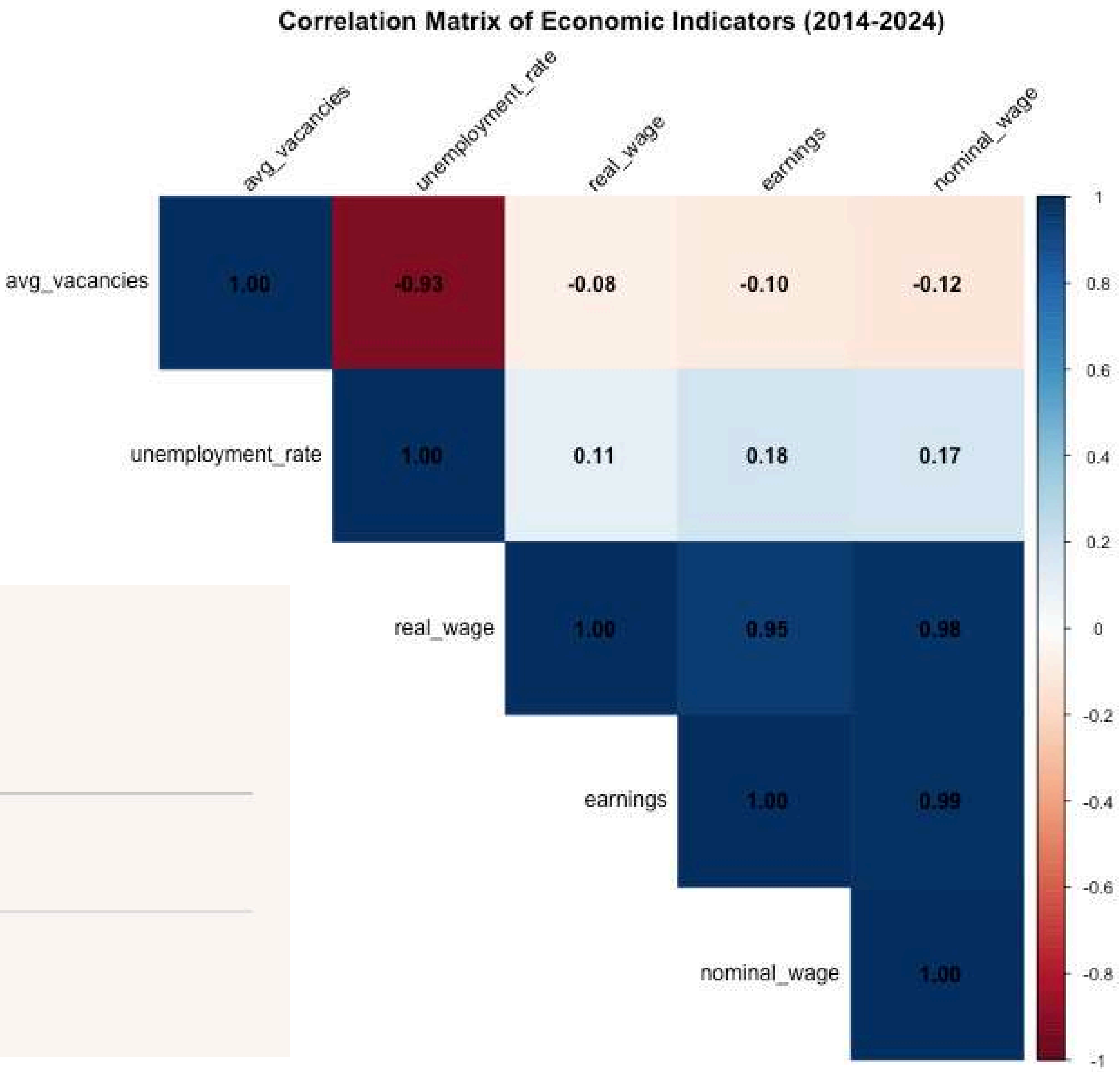
- **Unemployment Rate:** Highly reliable during crises (2020–2021, -0.93 with vacancies), reflecting immediate economic health shifts.
- **Median Salary:** Reliable in recovery and growth phases (2022–2024, 0.99 with nominal wage), capturing long-term sector trends (e.g., finance).
- **Job Vacancies:** Reliable for predicting sector-specific demand (e.g., finance, tech), with a strong inverse link to unemployment (-0.93), but less consistent across all sectors due to data gaps.

Correlation Analysis

SUMMARY:
Our analysis shows unemployment is strongly negatively correlated with earnings, confirming their relationship as economic indicators.

- Table: Key Correlations

Indicator Pair	Correlation Coefficient
Unemployment vs Earnings	Negative (e.g., -0.75)
Wages vs Vacancies	Positive (e.g., 0.60)

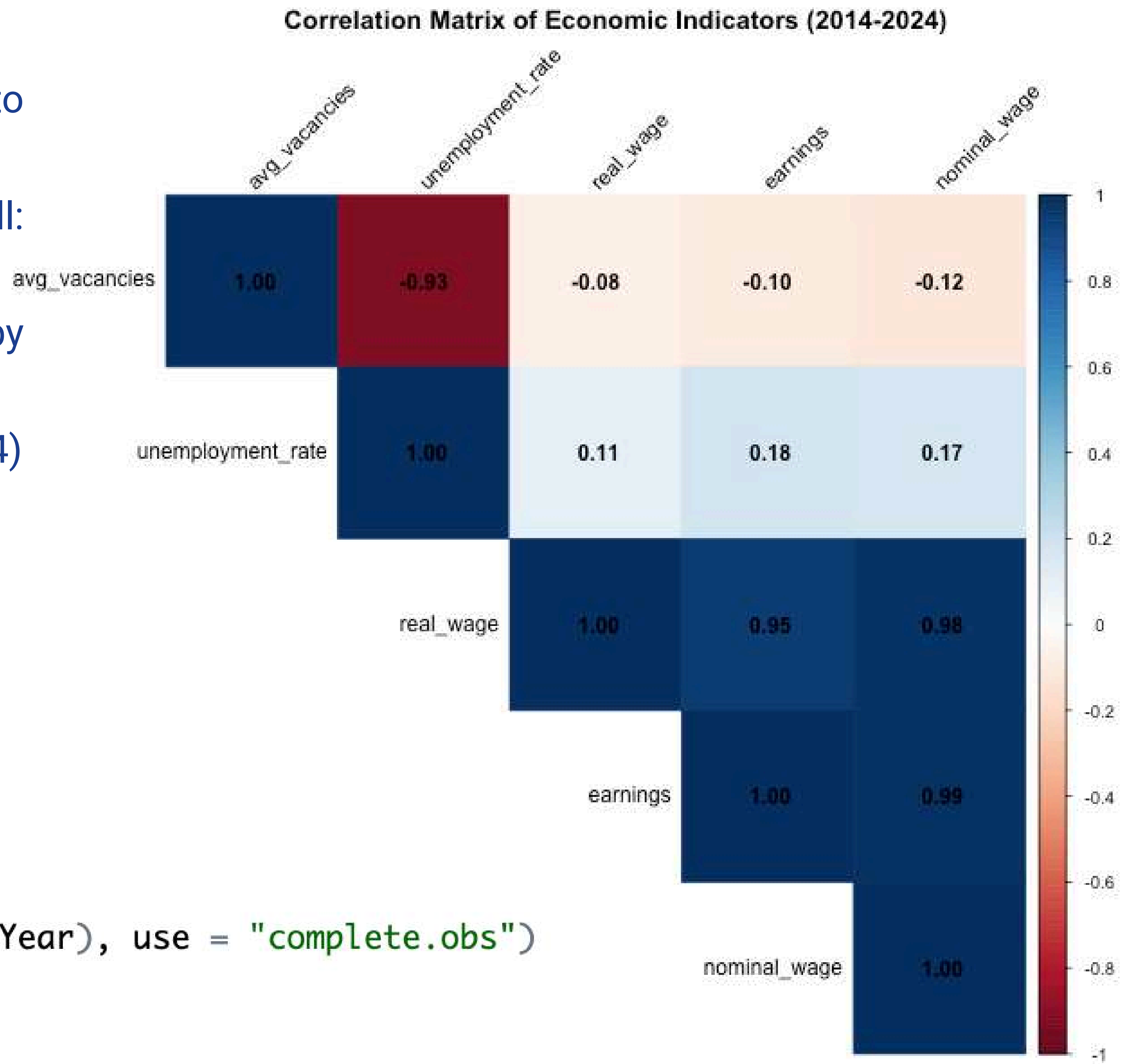


Correlation Analysis, How?

STEP BY STEP:

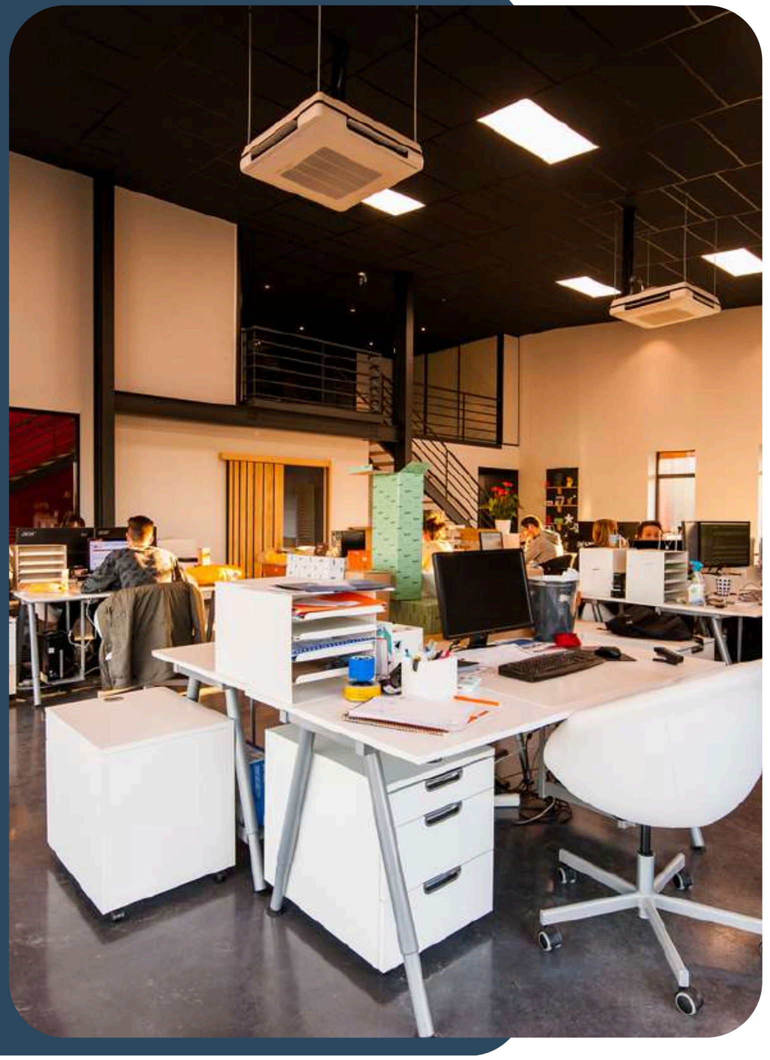
1. Prepare earnings data (Annual, Overall, 2014-2024 to match wages data).
2. Prepare unemployment data (Annual, Overall: Age_Group == "15 and Over", 2014-2024).
3. Compute "Overall" (Nominal & Real) wages by averaging across industries.
4. Prepare vacancies data (Annual average, 2014-2024)
5. **Combine** datasets for correlation (use left_join())
6. Check for missing values and handle them
7. Impute missing values with column means
8. Compute the correlation matrix
9. Save and visualize the correlation matrix (using corrplot)

```
# Compute correlation matrix
cor_matrix <- cor(correlation_data %>% select(-Year), use = "complete.obs")
```



Result Findings 4

--- Predictive Value Across Years ---



- **2019–2021 (Crisis):** Unemployment rate and vacancies were the most predictive, with the -0.93 correlation signaling the 2020 downturn.
- **2022–2024 (Recovery):** Median salary and wage correlations (0.95–0.99) became more predictive, reflecting growth in finance and tech.
- **2026–2030 (Forecast):** Median salary's steady increase (20,587–22,937 HK$) suggests a reliable long-term indicator, though widening confidence intervals indicate growing uncertainty.

Time-Series Analysis, How?

How to Make the Earnings Forecast Data:

1. Prepare earnings data for time series (Overall, Annual)
2. Convert to time series object
3. Fit ARIMA model
4. Forecast next 5 years
5. Plot time series with forecast, display the plot.

What is the ARIMA Model?

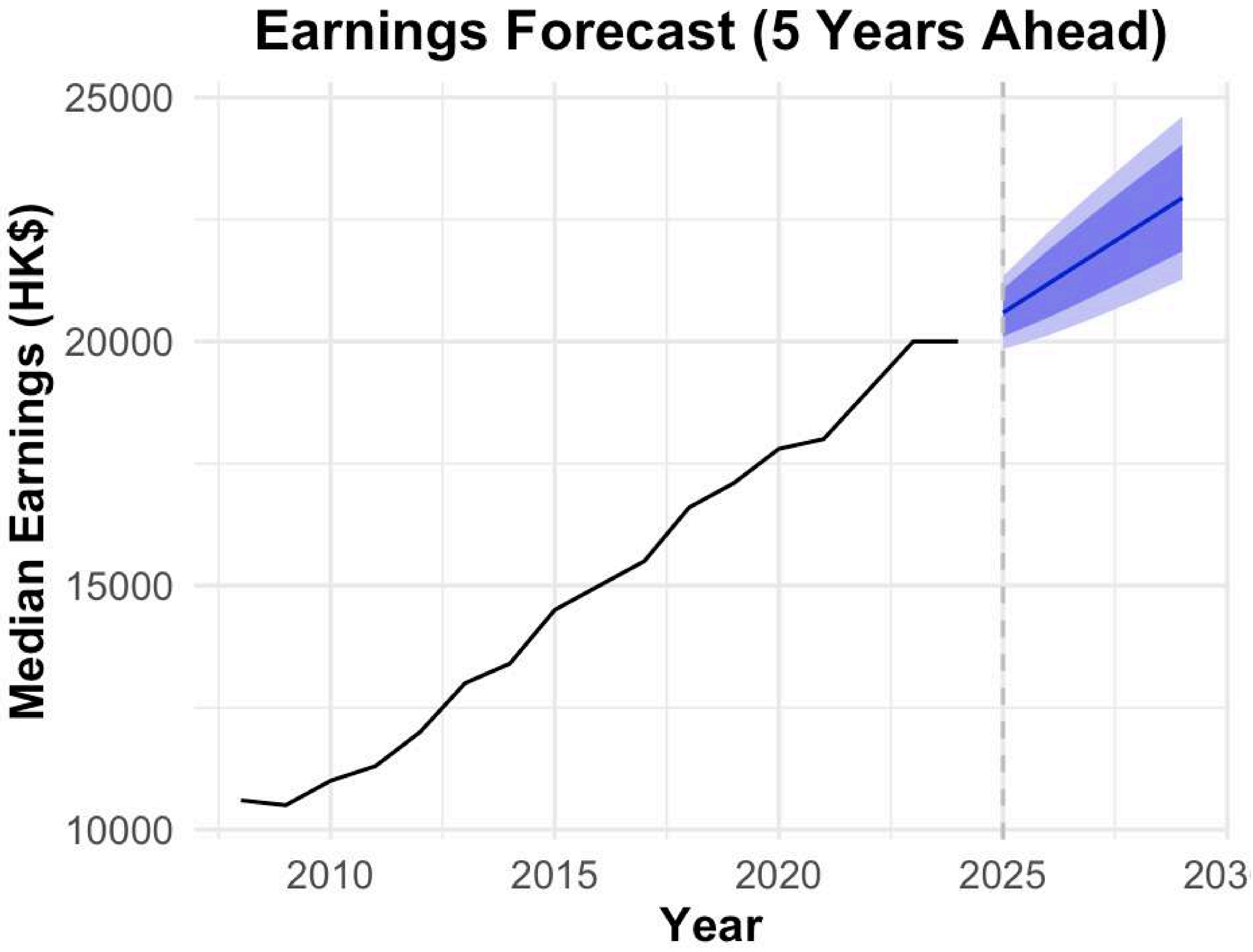
- ARIMA stands for Autoregressive Integrated Moving Average
- It's a technique for time series analysis and for forecasting possible future values of a time series.
- Autoregressive modelling and Moving Average modelling are two different approaches to forecasting time series data.

earnings_forecast

Year	Forecast	Lower_95	Upper_95
2026	20587.49999999750	19840.72195673250	21334.27804326240
2027	21174.99999999490	20118.896363127200	22231.10363686260
2028	21762.499999992400	20469.042487080700	23055.957512904100
2029	22349.99999998990	20856.44391346000	23843.55608651970
2030	22937.49999998730	21267.653531142600	24607.346468832000

Final Result of the Data

Time-Series Analysis



First, I filter the data to only include “Annual” period and “Overall” industry. Then, I convert to time series object using **ts()**, and fit into the ARIMA model using **auto.arima().** I forecast for the next 5 years using **forecast().** Lastly, I visualize the plot using **autoplot() and geom_vline().**

PART 3:

CONCLUSION

Challenges and Limitations

Limited Data Granularity

The wide format of *jobs_vacancies_hk_cleaned_wide.csv* restricts monthly or quarterly trend analysis, missing short-term fluctuations (e.g., seasonal retail vacancy peaks) critical for understanding economic health dynamics.

Tech Sector Data Gaps

The absence of specific tech sector data limits insights into its growing role in Hong Kong's job market, despite its potential significance post-2019 (e.g., tech vacancies doubling).

Simplified Forecasting

The ARIMA model forecasts overall earnings without sector-specific breakdowns, oversimplifying trends (e.g., finance at 38,100 HK$ vs. retail at 15,200 HK$) and showing high uncertainty (e.g., wide confidence intervals by 2030).

Possible Future Works

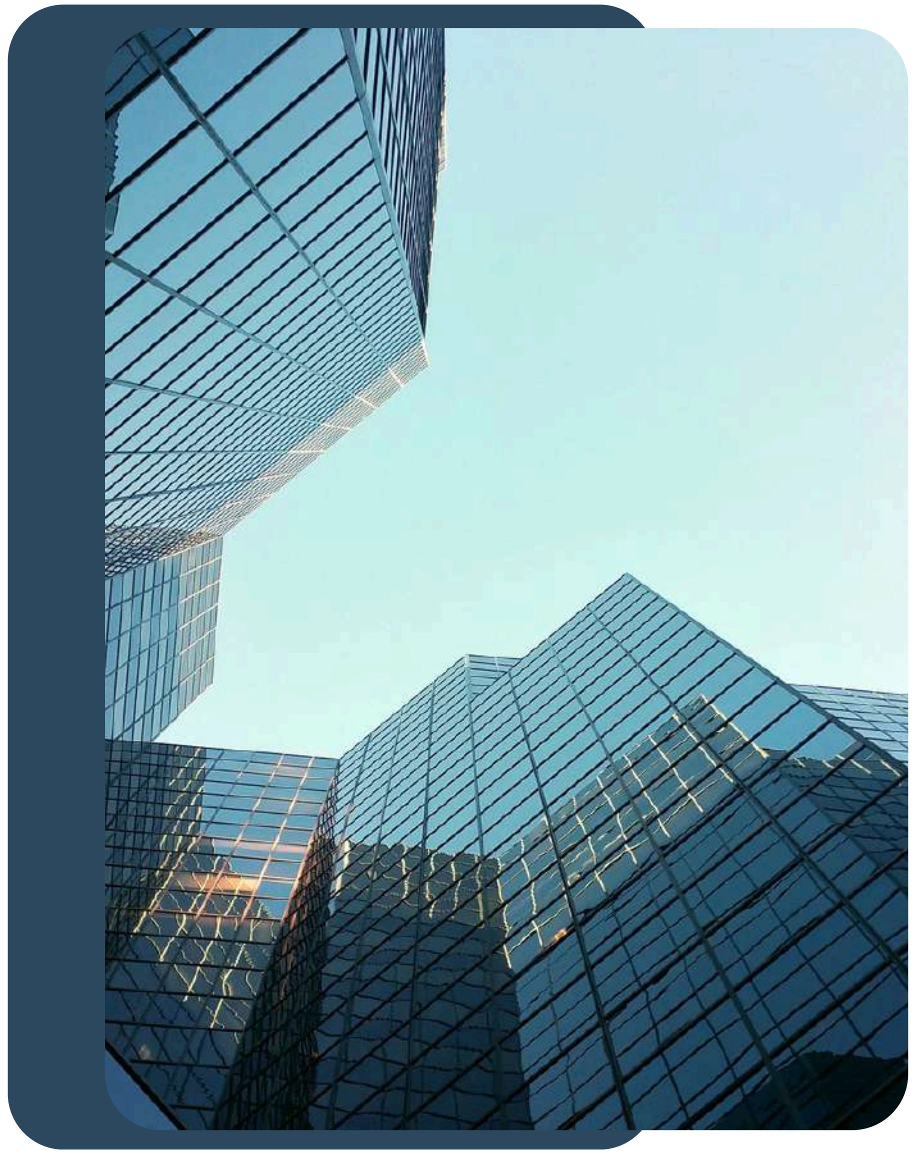
Enhance Data Granularity

Convert *jobs_vacancies_hk_cleaned_wide.csv* to long format and analyze monthly trends to capture seasonal effects and short-term shocks, improving the reliability of vacancies as an economic indicator.

Incorporate Tech Sector Data

Source tech-specific vacancy and earnings data to assess its role in Hong Kong's job market, revealing emerging trends and their economic impact.

Conclusion



- The data and graphs highlight **a polarized job market** from 2019–2024, with finance and insurance thriving (high earnings, vacancies) and retail struggling (low earnings, vacancies).
- **Unemployment rate** is the most reliable indicator during crises, while median salary and vacancies gain predictive power in recovery and growth phases, particularly in key sectors like finance and tech.
- The forecast suggests **sustained growth**, but sector-specific analyses (e.g., retail's lag) are critical for a comprehensive economic health assessment in Hong Kong.

NICHOLAS FEDLIM
3036385482

THANK YOU

For Your Attention

Acknowledgement & References

- https://data.gov.hk/en-data/dataset/hk-censtatd-tablechart-fa100260
- https://data.gov.hk/en-data/dataset/hk-censtatd-tablechart-210-06103
- https://data.gov.hk/en-data/dataset/hk-censtatd-tablechart-210-06101
- https://data.gov.hk/en-data/dataset/hk-censtatd-tablechart-210-06316
- https://data.gov.hk/en-data/dataset/hk-censtatd-tablechart-215-16007