Unemployment Study

Jaedin Hernandez

2025-09-09

Background

I will be conducting a study of the unemployment rate in major cities by following unemployment tends from 2019 to 2023. Our primary tools for research shall be ANOVA and multiple linear regression models to determine if unemployment rate is affected by factors such as region, age, % of unemployed with degrees, date, and job postings.

The primary use of this study will be to determine if there is a discrepancy in the unemployment factors year around, vs during months of college graduation. This is due to a high volume of persons being introduced into the work force during these periods, which could cause discrepancy in our overall results if not first targeted.

```
##Install SQL libraries
library(DBI)
library(RSQLite)
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
## Connect unemployment rate database to R
database <- dbConnect(SQLite(), "Unemployment.sqlite")

dbListTables(database)
```

```
[1] "GradMonthUnemployment" "MarketTrends"
```

```
dbListFields(database, "GradMonthUnemployment")
```

```
 [1] "id"                        "date"
 [3] "location"                  "unemployment_rate"
 [5] "job_postings"              "in_demand_skills"
 [7] "average_age"               "college_degree_percentage"
 [9] "year"                      "month"
```

Data prep for graduation period unemployment

```r
# Create a table for unemployment rates during graduation months
GradUnemployment <- tbl(database, "GradMonthUnemployment")
GradUnemployment
```

```
# Source:   table<`GradMonthUnemployment`> [?? x 10]
# Database: sqlite 3.50.4 [/Users/jaedin/Desktop/Tools for Data Science/Unemployment.sqlite]
      id date       location     unemployment_rate job_postings in_demand_skills
   <int> <chr>      <chr>                    <dbl>        <int> <chr>
 1     2 2025-05-24 Washington                11.1         2695 Data Analysis, ~
 2     4 2024-12-28 Indianapolis              11.2         3708 Cloud Computing~
 3     6 2024-12-31 Los Angeles                5           1785 Cloud Computing~
 4    12 2025-05-01 Washington                 2.3         2135 Project Managem~
 5    22 2025-05-10 Jacksonville               9.3         4254 Data Analysis, ~
 6    24 2024-12-31 Charlotte                  2           1892 Cybersecurity, ~
 7    28 2024-12-19 San Francis~               2.9         3376 Data Analysis, ~
 8    29 2024-06-25 Charlotte                  2.5         1377 Digital Marketi~
 9    32 2023-12-18 Jacksonville               5.2         3166 SQL, Machine Le~
10    36 2024-05-22 Dallas                    11.1          476 Cloud Computing~
# i more rows
# i 4 more variables: average_age <int>, college_degree_percentage <int>,
#   year <chr>, month <chr>
```

```r
GradUnemployment_df <- collect(GradUnemployment)

# Create Variables for each grad month column
Location_G <- GradUnemployment_df$location
Rate_G <- GradUnemployment_df$unemployment_rate
Postings_G <- GradUnemployment_df$job_postings
Age_G <- GradUnemployment_df$average_age
Degree_G <- GradUnemployment_df$college_degree_percentage
Year_G <- GradUnemployment_df$year
```

Calculations

We shall perform one way ANOVA for unemployment rate based on location for data collected during graduation months. (May, August, December)

```r
# Perform ANOVA for unemployment rate based on location in college graduation months
aovLocationRateG <- aov(Rate_G~Location_G, data=GradUnemployment_df)
aovLocationRateG
```

```
Call:
   aov(formula = Rate_G ~ Location_G, data = GradUnemployment_df)

Terms:
                Location_G Residuals
Sum of Squares     312.675  3260.715
Deg. of Freedom         19       229
```

```
Residual standard error: 3.773451
Estimated effects may be unbalanced
```

```r
SummaryLocationRateG <- summary(aovLocationRateG)
SummaryLocationRateG
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
Location_G   19    313   16.46   1.156  0.298
Residuals   229   3261   14.24
```

```r
## Perform multiple linear regression for unemployment in graduation months
RegressionModelG <- lm(Rate_G ~ Postings_G + Age_G + Degree_G + Year_G, data=GradUnemployment_df)
RegressionModelG
```

```
Call:
lm(formula = Rate_G ~ Postings_G + Age_G + Degree_G + Year_G,
    data = GradUnemployment_df)

Coefficients:
(Intercept)   Postings_G        Age_G     Degree_G   Year_G2024   Year_G2025
 10.7496812   -0.0001848   -0.0138354   -0.0118763   -0.8188544   -1.2096511
```

```r
SummaryStatsRegG <- summary(RegressionModelG)
SummaryStatsRegG
```

```
Call:
lm(formula = Rate_G ~ Postings_G + Age_G + Degree_G + Year_G,
    data = GradUnemployment_df)

Residuals:
   Min     1Q Median     3Q    Max
-6.477 -3.436  0.047  3.109  6.930

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.7496812  1.6234304   6.622 2.26e-10 ***
Postings_G  -0.0001848  0.0001729  -1.069   0.2863
Age_G       -0.0138354  0.0314146  -0.440   0.6600
Degree_G    -0.0118763  0.0140605  -0.845   0.3991
Year_G2024  -0.8188544  0.6277276  -1.304   0.1933
Year_G2025  -1.2096511  0.6789731  -1.782   0.0761 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.791 on 243 degrees of freedom
Multiple R-squared:  0.02257,   Adjusted R-squared:  0.002457
F-statistic: 1.122 on 5 and 243 DF,  p-value: 0.3491
```

Analysis

The results indicate that the predictors—job postings, age, degree percentage, and graduation year—are not statistically significant, with all p-values above 0.05 and an overall model p-value of 0.3491. Additionally, the model explains only 2.3% of the variance ($R^2 = 0.0226$), suggesting these variables are poor predictors of unemployment rate during graduation months.

Further calculations

Due to a lack of findings in our ANOVA test, we shall try using polynomial regression models for unemployment rate in graduation months based on job postings and college degree percentages.

```r
#Create Polynomial variable for job postings and degree percentage
Postings_G2 <- GradUnemployment_df$job_postings^2
Degree_G2 <- GradUnemployment_df$college_degree_percentage^2

#Run polynomial regression models for postings and degree in predicting rate
postings_poly <- lm(Rate_G ~ Postings_G + Postings_G2 + Age_G + Degree_G, data = GradUnemployment_df)
summary(postings_poly)
```

```
Call:
lm(formula = Rate_G ~ Postings_G + Postings_G2 + Age_G + Degree_G,
    data = GradUnemployment_df)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4875 -3.2960 -0.1472  3.2515  6.9203

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.017e+01  1.698e+00   5.991 7.45e-09 ***
Postings_G  -3.899e-04  7.080e-04  -0.551    0.582
Postings_G2  3.798e-08  1.389e-07   0.273    0.785
Age_G       -1.205e-02  3.163e-02  -0.381    0.703
Degree_G    -1.271e-02  1.410e-02  -0.901    0.368
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.808 on 244 degrees of freedom
Multiple R-squared:  0.009925,  Adjusted R-squared:  -0.006306
F-statistic: 0.6115 on 4 and 244 DF,  p-value: 0.6547
```

```r
degree_poly <- lm(Rate_G ~ Postings_G + Age_G + Degree_G + Degree_G2, data = GradUnemployment_df)
summary(degree_poly)
```

```
Call:
lm(formula = Rate_G ~ Postings_G + Age_G + Degree_G + Degree_G2,
    data = GradUnemployment_df)

Residuals:
```

```
    Min      1Q  Median      3Q     Max
-6.6820 -3.3171 -0.2217  3.2554  7.0751

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.5110475  3.4492363   2.178   0.0304 *
Postings_G  -0.0001974  0.0001733  -1.139   0.2558
Age_G       -0.0096341  0.0315485  -0.305   0.7603
Degree_G     0.0730356  0.1071842   0.681   0.4963
Degree_G2   -0.0007049  0.0008722  -0.808   0.4197
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.803 on 244 degrees of freedom
Multiple R-squared:  0.01227,   Adjusted R-squared:  -0.003926
F-statistic: 0.7575 on 4 and 244 DF,  p-value: 0.5539
```

Data prep for unemployment rates(Year Around)

```
# Create table for overall unemployment rates
Unemployment <- tbl(database, "MarketTrends")
Unemployment
```

```
# Source:   table<`MarketTrends`> [?? x 10]
# Database: sqlite 3.50.4 [/Users/jaedin/Desktop/Tools for Data Science/Unemployment.sqlite]
      id date       location     unemployment_rate job_postings in_demand_skills
   <int> <chr>      <chr>                    <dbl>        <int> <chr>
 1     1 2023-10-07 Houston                    6.8         4894 Agile Methodolo~
 2     2 2025-05-24 Washington                11.1         2695 Data Analysis, ~
 3     3 2024-09-28 Chicago                    7.3         1174 Agile Methodolo~
 4     4 2024-12-28 Indianapolis              11.2         3708 Cloud Computing~
 5     5 2023-09-10 New York                  13.7          268 SQL, Machine Le~
 6     6 2024-12-31 Los Angeles                5           1785 Cloud Computing~
 7     7 2023-09-01 Phoenix                   13.7         2784 Data Analysis, ~
 8     8 2024-10-08 San Jose                  10.1         4981 Customer Servic~
 9     9 2024-08-06 Austin                     9           2453 Digital Marketi~
10    10 2024-02-14 Dallas                    12.4          808 Cloud Computing~
# i more rows
# i 4 more variables: average_age <int>, college_degree_percentage <int>,
#   year <chr>, month <chr>
```

```
Unemployment_df <- collect(Unemployment)
summary(Unemployment_df)
```

```
       id              date             location         unemployment_rate
 Min.   :   1.0   Length:1000        Length:1000        Min.   : 2.00
 1st Qu.: 250.8   Class :character   Class :character   1st Qu.: 5.40
 Median : 500.5   Mode  :character   Mode  :character   Median : 8.80
 Mean   : 500.5                                         Mean   : 8.63
 3rd Qu.: 750.2                                         3rd Qu.:11.80
 Max.   :1000.0                                         Max.   :15.00
```

```
  job_postings  in_demand_skills    average_age    college_degree_percentage
 Min.   :  53   Length:1000        Min.   :25.00   Min.   :30.00            
 1st Qu.:1213   Class :character   1st Qu.:31.00   1st Qu.:46.00            
 Median :2498   Mode  :character   Median :38.00   Median :60.00            
 Mean   :2495                      Mean   :37.86   Mean   :60.61            
 3rd Qu.:3779                      3rd Qu.:44.00   3rd Qu.:75.00            
 Max.   :4997                      Max.   :50.00   Max.   :90.00            
     year              month          
 Length:1000        Length:1000       
 Class :character   Class :character  
 Mode  :character   Mode  :character  
```

```r
# Create variables for each unemployment column
Location_U <- Unemployment_df$location
Rate_U <- Unemployment_df$unemployment_rate
Postings_U <- Unemployment_df$job_postings
Skills_U <- Unemployment_df$in_demand_skills
Degree_U <- Unemployment_df$college_degree_percentage
Age_U <- Unemployment_df$average_age
```

Calculations We shall perform one way ANOVA for unemployment rates

```r
aovLocationRate <- aov(Rate_U~Location_U, data=Unemployment_df)
aovLocationRate
```

```
Call:
   aov(formula = Rate_U ~ Location_U, data = Unemployment_df)

Terms:
                Location_U Residuals
Sum of Squares     193.921 13513.987
Deg. of Freedom         19       980

Residual standard error: 3.71346
Estimated effects may be unbalanced
```

```r
SummaryStatsLocationRate <- summary(aovLocationRate)
SummaryStatsLocationRate
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
Location_U   19    194   10.21    0.74  0.779
Residuals   980  13514   13.79               
```

Analysis

We notice that our P-Value(0.779)>alpha(0.05) thus we do not reject H0 as the mean unemployment rate does not differ among different cities in this dataset. Theirfore we shall run a multiple linear regression model to determine if any numerical predictors appear to be a good fit.

Regression model

```
RegressionModel_U <- lm(Rate_U ~ Postings_U + Degree_U + Age_U, data=Unemployment_df)
RegressionModel_U
```

```
Call:
lm(formula = Rate_U ~ Postings_U + Degree_U + Age_U, data = Unemployment_df)

Coefficients:
(Intercept)   Postings_U     Degree_U        Age_U
  8.634e+00    5.259e-05   -5.289e-03    4.894e-03
```

```
SummaryReg <- summary(RegressionModel_U)
SummaryReg
```

```
Call:
lm(formula = Rate_U ~ Postings_U + Degree_U + Age_U, data = Unemployment_df)

Residuals:
   Min     1Q Median     3Q    Max
-6.760 -3.208  0.163  3.158  6.605

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.634e+00  7.521e-01  11.479   <2e-16 ***
Postings_U   5.259e-05  8.171e-05   0.644    0.520
Degree_U    -5.289e-03  6.754e-03  -0.783    0.434
Age_U        4.894e-03  1.542e-02   0.317    0.751
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.708 on 996 degrees of freedom
Multiple R-squared:  0.001098,  Adjusted R-squared:  -0.001911
F-statistic: 0.3649 on 3 and 996 DF,  p-value: 0.7783
```

Analysis

We notice that our numerical predictors (degree%, Postings, and Age) do not have a p-value<0.05, theirfore our numerical predictors are determined to have little effect on unemployment rate in this data set. Thus we shall attempt to find discrepencies in rate by catagorical data.

Plots for unemployment by month

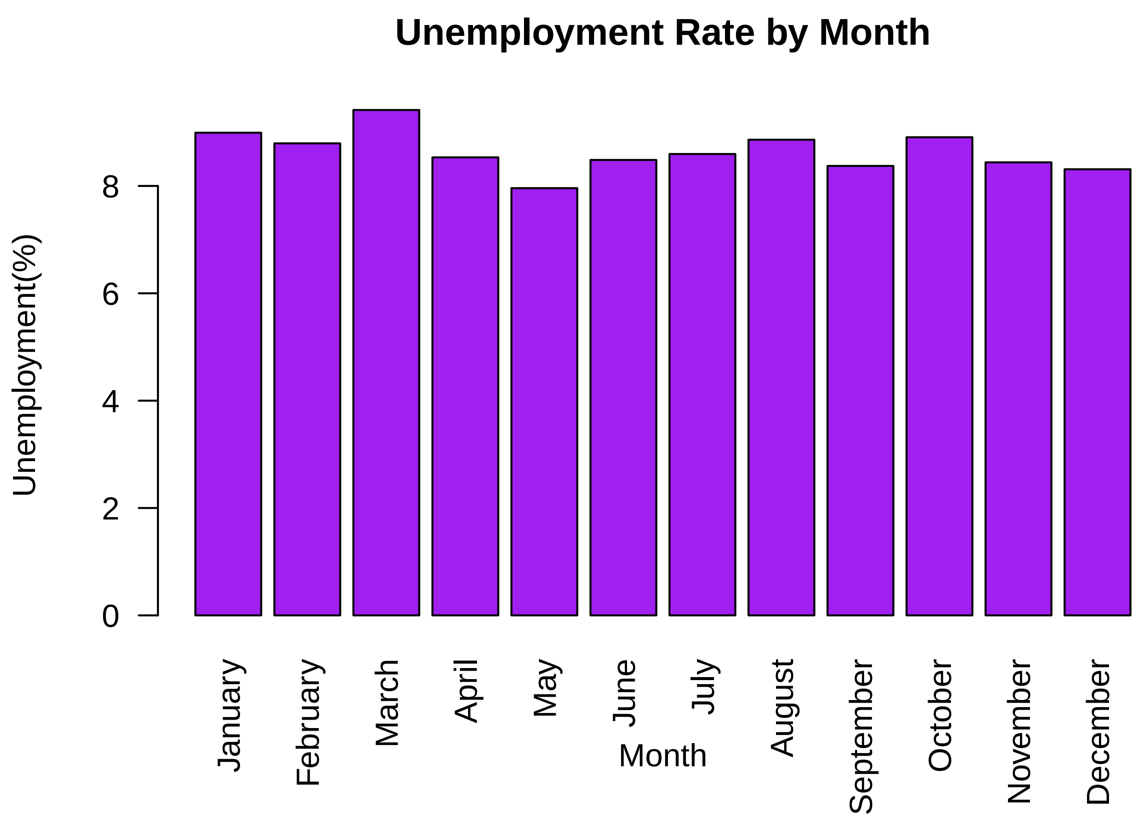
```
#Convert date column to Date format
Unemployment_df$date <- as.Date(Unemployment_df$date)

#Extract month
```

```r
Unemployment_df$Month <- format(Unemployment_df$date, "%B")
Unemployment_df$Month <- factor(Unemployment_df$Month,
                                levels = month.name)

#Average Unemployment rate by month
avg_rate_by_month <- tapply(Unemployment_df$unemployment_rate, Unemployment_df$Month, mean, na.rm = TRU

barplot(avg_rate_by_month,
        main= "Unemployment Rate by Month",
        xlab= "Month",
        ylab= "Unemployment(%)",
        col= "purple",
        las = 2)
```

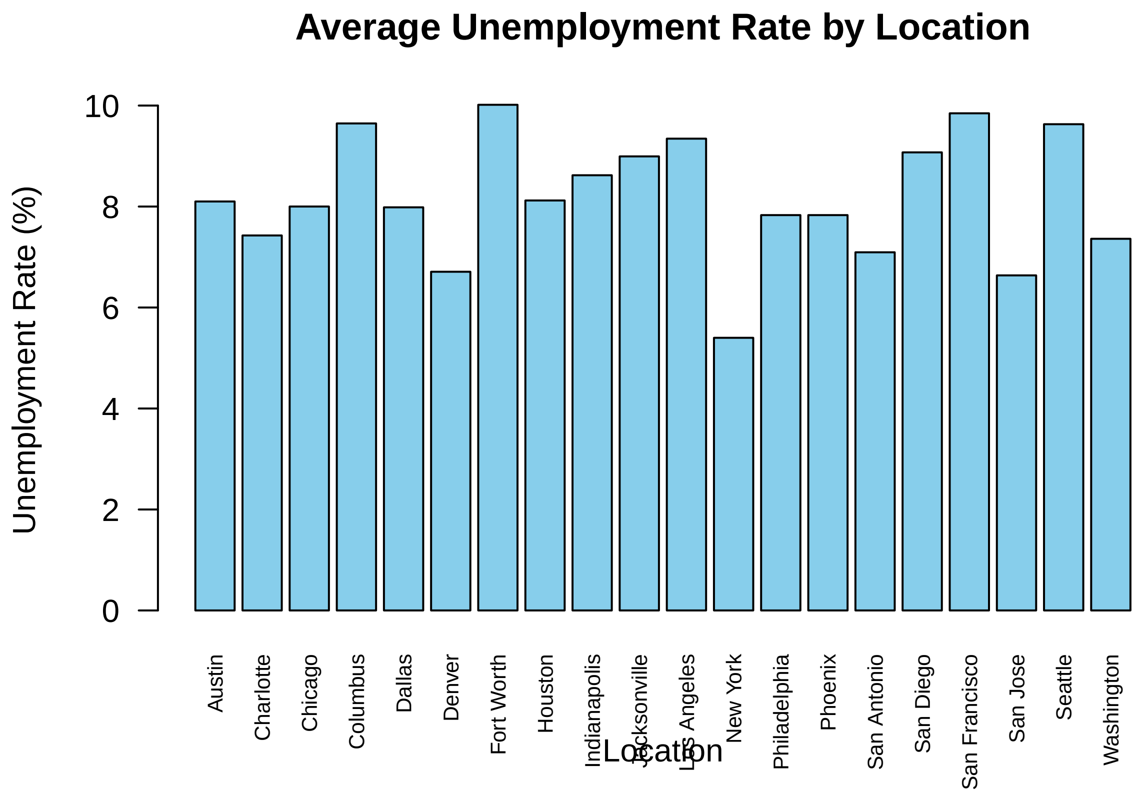


Plots for unemployment by location(Year Around)

```r
# Calculate mean unemployment rate per location
avg_rate_by_location <- tapply(Rate_G, Location_G, mean, na.rm = TRUE)

# Create bar plot
barplot(avg_rate_by_location,
        main = "Average Unemployment Rate by Location",
        xlab = "Location",
```

```
        ylab = "Unemployment Rate (%)",
        col = "skyblue",
        las = 2,           # Rotate x-axis labels
        cex.names = 0.7)  # Shrink label size if names are long
```



Analysis

We notice New York has the lowest unemployment rate by location, theirfore we should try to target this area in our study in order to discover any discrepancys in the time period or job postings.

We also determine that Fort Worth is the leader in unemployment regions in this study. For this we shal further analyze the unemployment rates over time for this region.

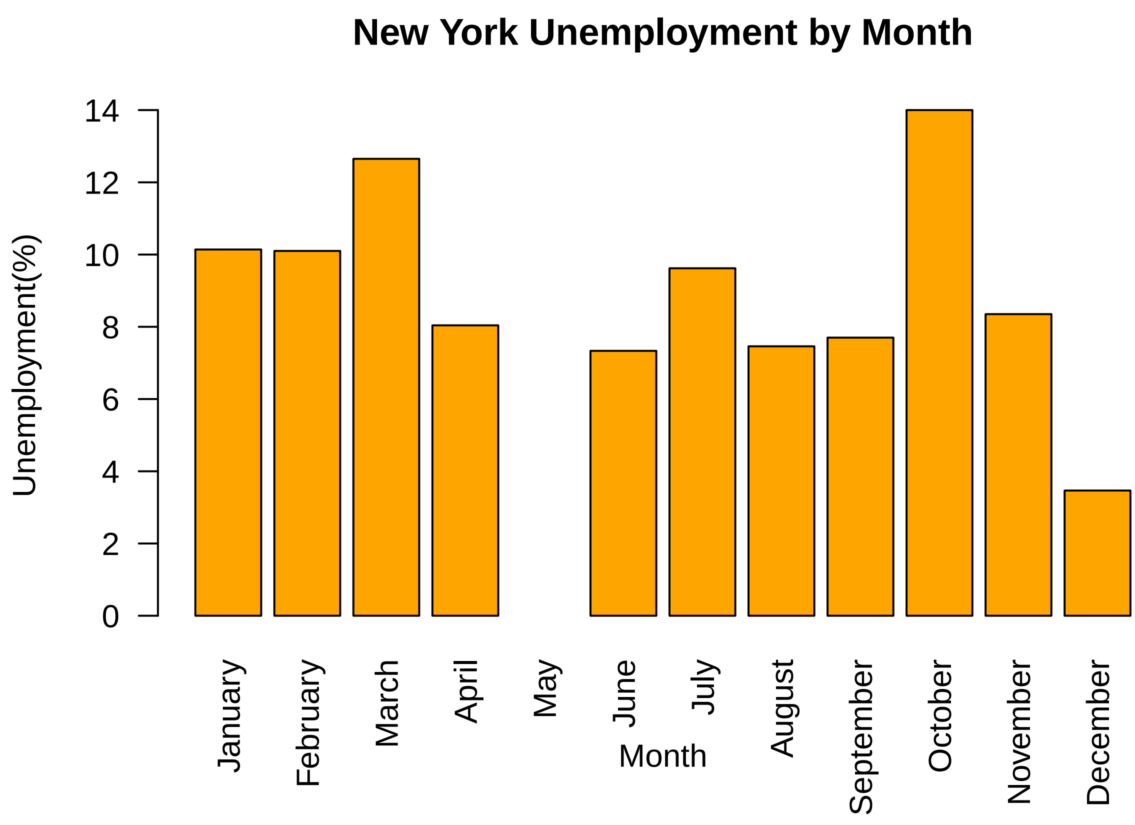
Unemployment rate by month in New York

```
# Filter data set for New York location only
Ny_Unemployment <- subset(Unemployment_df, location == "New York")

#Convert date column to Date format
Ny_Unemployment$date <- as.Date(Ny_Unemployment$date)

#Extract month
Ny_Unemployment$Month <- format(Ny_Unemployment$date, "%B")
```

```r
Ny_Unemployment$Month <- factor(Ny_Unemployment$Month,
                                    levels = month.name)
#Calculate average unemployment by month in New York
NyRate_by_month <- tapply(Ny_Unemployment$unemployment_rate,
                          Ny_Unemployment$Month,
                          mean,
                          na.rm=TRUE)
barplot(NyRate_by_month,
        main = "New York Unemployment by Month",
        xlab="Month",
        ylab="Unemployment(%)",
        col="orange",
        las=2)
```



Analysis

We notice by this graph, unemployment appears to spike during the months of March, July, and October. These also happen to be the months where quarterly reports are released for most large corporations. This, along with other factors could lead to a discrepency in unemployment, rather than periods of college graduation. Further study targeting these months, along with corporate employment data may be required for further analysis.

Unemplyment rate by month in Fort Worth

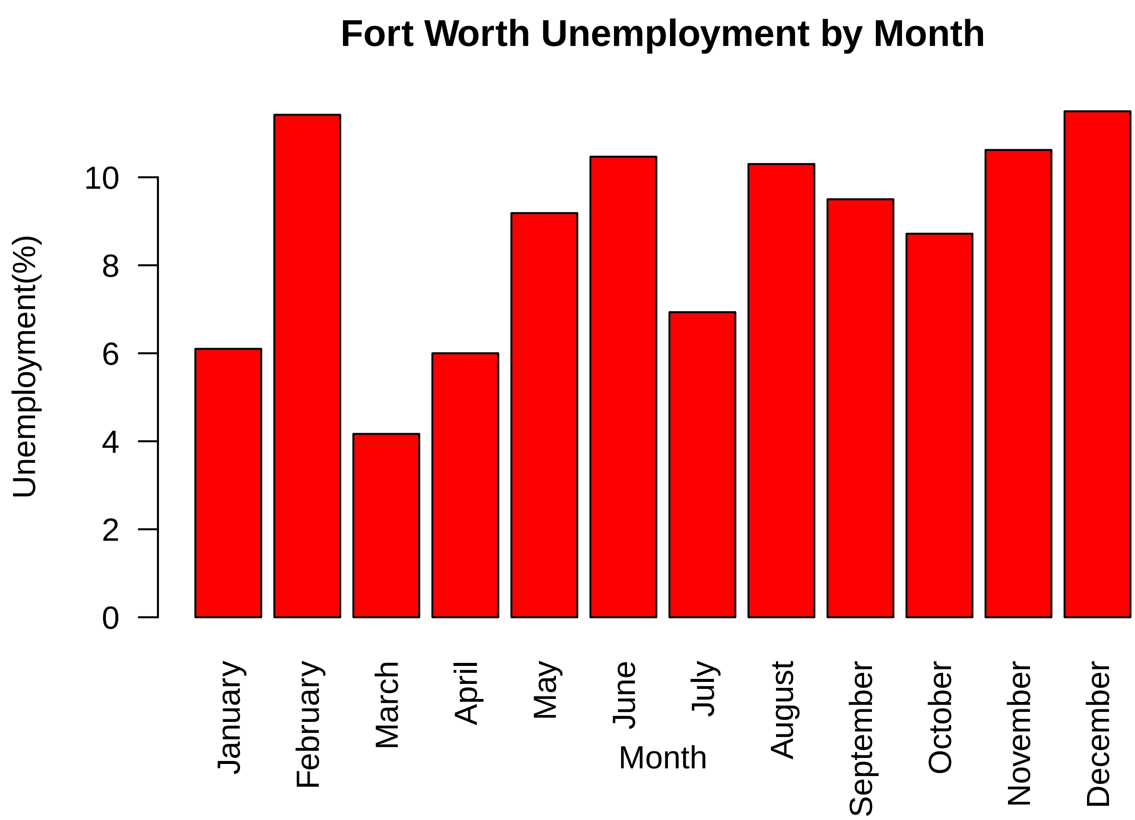
```
FW_Unemployment <- subset(Unemployment_df, location == 'Fort Worth')

FW_Unemployment$date <- as.Date(FW_Unemployment$date)

FW_Unemployment$Month <- format(FW_Unemployment$date, "%B")
FW_Unemployment$Month <- factor(FW_Unemployment$Month,
                                levels = month.name)

FW_unemp_by_month <- tapply(FW_Unemployment$unemployment_rate,
                            FW_Unemployment$Month, mean,
                            na.rm = TRUE)

barplot(FW_unemp_by_month,
        main = "Fort Worth Unemployment by Month",
        xlab = "Month",
        ylab = "Unemployment(%)",
        col = "red",
        las = 2)
```



Analysis

The unemployment rate in Fort Worth peaks notably in February, May, and August. This likely reflects the expiration of many military service contracts in December and January, resulting in a surge of civilian career transitions in February, a pattern common in military base regions. Additionally, May and August align with major college graduation months, contributing to higher unemployment rates. December also shows elevated unemployment, possibly due to a combination of expiring contracts and fall graduates entering the workforce.

Conclusion

In this study, we applied ANOVA and multiple linear regression to examine the relationship between unemployment rate and various factors including time of year, region, average age, job postings, and education level. Our statistical models showed no significant correlation between unemployment rate and numerical predictors, with p-values above 0.05 and low R^2 values.

However, visual analysis revealed notable patterns. Cities like Fort Worth and New York exhibited unemployment spikes during months tied to college graduations, military contract expirations, and corporate layoffs. These trends suggest that while our models did not detect strong statistical relationships, categorical and seasonal factors may still play a significant role in unemployment variation and merit further investigation.