Analysis of Annual Compensation in Information Technology

CIS663-01-SP24 Term Project Report

Mark M. Young (M00348726)

May 11, 2024

Table of Contents

[1 Abstract 2](#_Toc166359271)

[2 Introduction 2](#_Toc166359272)

[3 Literature Review 2](#_Toc166359273)

[4 Theory 3](#_Toc166359274)

[5 Data 3](#_Toc166359275)

[5.1 File Importing 3](#_Toc166359276)

[5.2 Variable Selection 3](#_Toc166359277)

[5.3 Variable Splitting 4](#_Toc166359278)

[5.4 Variable Conversion 4](#_Toc166359279)

[5.4.1 Numeric Variables 4](#_Toc166359280)

[5.4.2 Categorical Variables 4](#_Toc166359281)

[5.4.3 Ordinal Variables 5](#_Toc166359282)

[5.5 Value Filtering 5](#_Toc166359283)

[6 Methodology 6](#_Toc166359284)

[6.1 Shape of the Data 6](#_Toc166359285)

[6.2 Normality of the Data 12](#_Toc166359286)

[6.3 Kurtosis and Skewness of the Data 14](#_Toc166359287)

[7 Results 14](#_Toc166359288)

[7.1 Simple Linear Regressions 14](#_Toc166359289)

[7.2 Multiple Linear Regression 17](#_Toc166359290)

[7.3 Interpretation 20](#_Toc166359291)

[7.4 Example Prediction 21](#_Toc166359292)

[8 Implications 21](#_Toc166359293)

[9 Conclusion 22](#_Toc166359294)

[10 References 22](#_Toc166359295)

# 1 Abstract

This research applies linear regression and graphical analysis to data from the 2022 StackOverflow.com Developer Survey, highlighting the significance of developer type, years of experience, and gender in information technology (IT). Using voluntary responses from over 70,000 participants, insights were made into the association with the dependent variable annual compensation and among the independent variables developer’s age, education level, years of experience, gender, organization size, position type, and remote work, with some used as control variables. This report explains data, analysis, methods, and results from a graduate, educational setting. The results are that female IT professionals do have a lower median income, but are in fact compensated $2,672.13 more per year, holding all other variables constant.

# 2 Introduction

Compensation in IT is a widely discussed topic from multiple aspects. Economists discuss the implications of onshore (domestic), offshore, and nearshore development (Nguyen, 2022); managers discuss the cost of IT due to the ubiquity of IT in conducting business (Bajarin, 2020); and professionals discuss job opportunities during periods of decreased IT demand, like in 2002 (Shim, 2003), 2009 (Ballenstedt, 2010), and 2023 (Inspirisys, 2024). This research focuses on the implications from the IT professionals’ perspective.

IT professionals must adapt to the changes and advancements in IT or their compensation will stagnate. The aim of this research is to expose statistically significant variables which influence compensation. With the results, IT professionals can make informed decisions when considering professional development, seeking new employment, or interested in changing to a different position type.

# 3 Literature Review

A generalized investigation was performed by Coursera.org using data from Dice.com, GlassDoor.com, et al. in March of 2024. It revealed a wide range of compensation depending on living location and position type. It found that simply moving from Michigan to California could increase an IT professional’s salary by over 50%, but did not take into account the difference in the cost of living index of the respective locations (Coursera, 2024).

Specifically to the variable of gender, many articles discuss how females make 16% less than males. However, these articles merely group females and males by industry without considering, or at least listing, job role, years of experience, or education level (Haan, 2024). One article discussed how female nurses received less yearly compensation than males, but went on to admit that males were 29% more likely to negotiate a higher salary, worked 5.4% more regular hours, and 25% more overtime hours (Carlow, 2022).

Specifically to the professionals in IT, the Information Week Salary Survey of 2022 claimed that females received 82% the yearly compensation of what males received. This survey seemed to a survey which specifically set out to ask 1) what is your gender, 2) how much do you make, and 3) are you in management (Davis, 2022). The question became, “What would the results be if those questions were answered as part of larger survey?”

# 4 Theory

Since World War II, especially since 1985, women are more likely to work in once exclusively male occupations and industries and vice versa (Nichols & Zimmerman, 2008). This has helped reduce household income volatility, but has introduced controversy on a larger scale with one side of the gender lines claiming that another side is being paid more, or claiming that one side has lower expectation of productivity. This study’s aim is to analyze the voluntary survey results of over 70,000 respondents to determine if there is a statistically significant difference between genders, and discover other contributing variables using linear regression analysis.

:

:

# 5 Data

The dataset available from StackOverflow.com is in CSV (Comma-Separated Value) format, contains 79 variables, is over 100 MB (megabytes), and can be viewed using a spreadsheet application like Microsoft Excel. However, CSV data is a string of characters at its basis with no further meaning (e.g., 1.618 is a string of five characters, not a number). Technically, R does not even assume the character strings can be legitimately interpreted as character data. Furthermore, most of the variables of interest were answered in the survey from a list of options and some of those options represent a range of values. Further still, a few values have a meaningful order to them. For this reason, the values were converted for interpretability when plotting and performing numerical analysis. The data is imported using the built-in R function read.csv():

## 5.1 File Importing

dat\_file\_name <- "stack-overflow-developer-survey-2022.survey\_results\_public.csv"  
dat\_raw <- read.csv(dat\_file\_name)

## 5.2 Variable Selection

Although some variables did not play an eventual role in the results, a preliminary selection of variables was performed to reduce the data to twelve variables:

keep\_column\_names <- c(  
 "Age", "ConvertedCompYearly", "Country", "DevType",   
 "EdLevel", "Employment", "Ethnicity", "Gender",  
 "OrgSize", "RemoteWork", "ResponseId", "YearsCodePro"  
)  
# Only keep columns of interest.  
dat\_selected <- dat\_raw[keep\_column\_names]

## 5.3 Variable Splitting

The DevType variable is a composite variable with values delimited by a semicolon (;). To effectively analyze this variable, it needs to be split into multiple rows using separate\_longer\_delim (Wickham, Vaughan, & Girlich, 2024).

# Split.  
 dat\_selected <- dat\_selected %>% separate\_longer\_delim(c(DevType), delim=";")

Replacement of long value labels omitted for brevity.

# Transform.  
 dat\_selected$DevType.f <- factor(dat\_selected$DevType,  
 levels=dev\_type\_cnt$x[order(dev\_type\_cnt$freq, decreasing=FALSE)]  
 )

## 5.4 Variable Conversion

### 5.4.1 Numeric Variables

The dependent variable for this analysis is annual compensation. The ConvertedCompYearly column is used for this value because the CompFreq, CompTotal, Country, and Currency columns would have be considered collectively with a specific date in order to convert the values to a common currency for comparison. The values are explicitly converted to numeric values:

# Transform numerical variables.  
dat\_selected$ConvertedCompYearly.n <- as.numeric(as.character(dat\_selected$ConvertedCompYearly))

dat\_selected$YearsCodePro.n <- as.numeric(as.character(dat\_selected$YearsCodePro))

### 5.4.2 Categorical Variables

This study is going to specifically focus on the compensation of Females versus all other Genders. Therefore, all non-Female, non-NA values are replaced with Other. [The raw dataset contained the value of “Woman” so this is replaced first.]

# Copy variable.  
dat\_selected$Gender.d <- dat\_selected$Gender  
# Replace all values that are not "Woman", or NA with "Other".  
dat\_selected$Gender.d[!dat\_selected$Gender %in% c("Woman", NA)] <- "Other"  
# Assign dummy values.  
dat\_selected$Gender.d[dat\_selected$Gender.d == "Other"] <- 0  
dat\_selected$Gender.d[dat\_selected$Gender.d == "Woman"] <- 1  
# Explicitly convert.  
dat\_selected$Gender.d <- as.integer(dat\_selected$Gender.d)  
dat\_selected$Gender.d <- factor(  
 dat\_selected$Gender.d,  
 levels=c(0, 1),  
 labels=c("Other", "Female")  
)

Now that DevType has been split, R needs to be explicitly told that it can now consider the variable categorical using the factor function.

# Transform.  
dat\_selected$DevType.f <- factor(dat\_selected$DevType)

### 5.4.3 Ordinal Variables

The OrgSize is a positive econometric value which needs to be interpreted by its logarithmic (or just log) value so change in organization size will scale correctly (Wooldridge, 2008, p. 46).

# Copy variable.  
dat\_selected$OrgSize.log <- dat\_selected$OrgSize

Replacement of character values with integer values omitted for brevity.

# Copy to logarithmic variable.  
dat\_selected$OrgSize.log <- log(dat\_selected$OrgSize.log)

## 5.5 Value Filtering

For the remainder of statistical analysis, only observations with provided ConvertedCompYearly.n, DevType.f, Gender.d, OrgSize.log, and YearsCodePro.n values will be considered.

# Remove observations without values in the variables in interest.  
dat\_sliced <- dat\_selected[  
 !is.na(dat\_selected$ConvertedCompYearly.n)  
 & !is.na(dat\_selected$DevType.f)  
 & !is.na(dat\_selected$Gender.d)  
 & !is.na(dat\_selected$OrgSize.log)  
 & !is.na(dat\_selected$YearsCodePro.n)  
 # Be sure to specify an comma at the end to avoid "undefined columns selected" error.  
 ,  
]

Additionally, considerable outliers were found within the raw data with IT professionals who made in excess of US$20M in 2022. Retaining these values makes graphs more difficult to interpret. For that reason, only observations with a yearly compensation less than US$250,000 will be considered.

# Remove outliers.  
dat\_sliced <- dat\_sliced[dat\_sliced$ConvertedCompYearly.n < 250000,]

# 6 Methodology

Initially, the study was going to use backward stepwise elimination, but the first run took 2 hours, 36 minutes to run and produced a 20 MB linear regression summary table (in HTML). For efficiency, forward stepwise inclusion was used when confirming the descriptive statistics (p-value, adj. R^2, and F-stat.) of variable in subsequent regressions. The p-values produced in the initial regression were used to prioritize subsequent forward stepwise inclusion of variables.

## 6.1 Shape of the Data

Using ggdensity (Kassambara, 2023), a density plot of the Yearly Compensation reveals an expected distribution showing that some, but few IT professionals make much more than others. There are spikes in density at the US$120k, US$150k, US$180k, and US$200k intervals. It is suspected that this is the result of survey participants rounding their compensation to the nearest US$5k (Figure 6-1).

ggdensity(dat\_sliced$ConvertedCompYearly.n,  
 main="Yearly Compensation",  
 ylab="Density",  
 xlab="Compensation (US$)"  
) +  
geom\_vline(aes(xintercept=median(dat\_sliced$ConvertedCompYearly.n)),  
 colour="darkgray", linetype="dashed"  
)

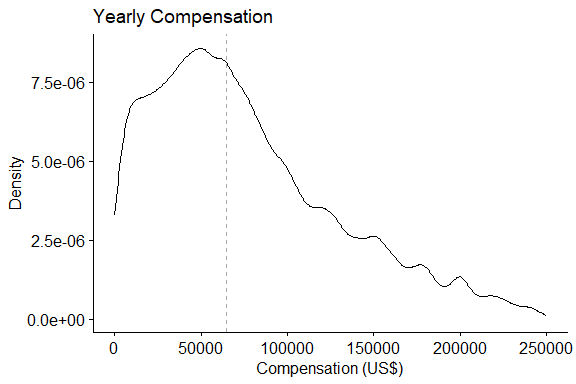


Figure 6-1: Yearly Compensation Density Plot

A density plot of the Years Coding Professionally reveals an expected distribution showing that there are more young IT professionals than older ones since high-technology is becoming more ubiquitous. There are spikes in density at the 10, 15, 20, 25, 30, 35, and 40 year intervals. It is suspected that this is the result of survey participants rounding their years of professional coding to the nearest 5 years (Figure 6-2).

ggdensity(dat\_sliced$YearsCodePro.n,  
 main="Years Coding Professionally",  
 ylab="Density",  
 xlab="Years Coding"  
) +  
geom\_vline(aes(xintercept=median(dat\_sliced$YearsCodePro.n)),  
 colour="darkgray", linetype="dashed"  
)

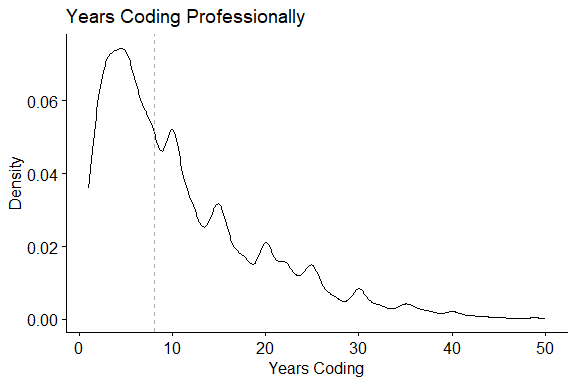


Figure 6-2: Years Code Pro. Density Plot

A bar chart of Gender reveals the just how much non-Females outnumber Females in IT professional positions (Figure 6-3).

f.gender <- barplot(  
 table(dat\_sliced$Gender.d),  
 main="Gender Frequency",  
 ylab="Count",  
 xlab="Gender"  
)

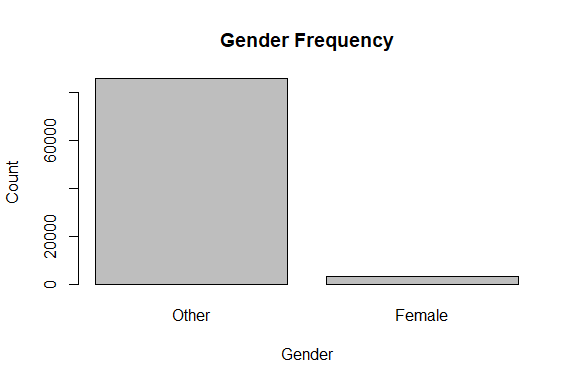


Figure 6-3: Gender Frequency Bar Chart

A density chart appears a bit like stalagmites. This is an artifact of the logarithm of the organization size coming from the arithmetic mean of the bounds of the OrgSize in the raw data (e.g., “20 to 99 employees” became ). So, although the values are numeric, they are limited to the original nine values (Figure 6-4).

ggdensity(dat\_sliced$OrgSize.log,  
 main="Organization Size Frequency",  
 ylab="Density",  
 xlab="Organization Size"  
) +  
geom\_vline(aes(xintercept=median(dat\_sliced$OrgSize.log)),  
 colour="darkgray", linetype="dashed"  
)

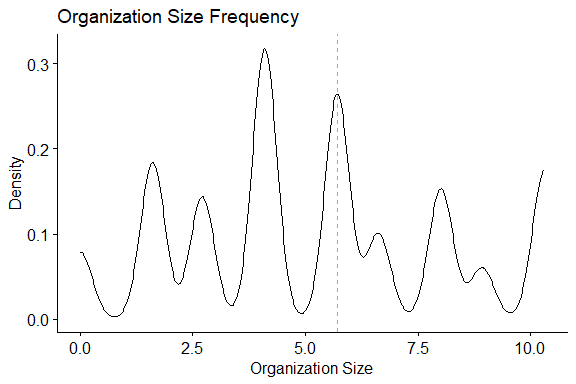


Figure 6-4: Organization Size Frequency Density Plot

A bar chart of reveals a frequency of Developer Types being mostly centered around Developer, full-stack; Developer, back-end; Developer, front-end; and Developer, applications (Figure 6-5).

par(mar=c(5.1, max(4.1, 17/1.8 ), 4.1, 2.1))  
f.devtype <- barplot(  
 table(dat\_sliced$DevType.f),  
 main="Dev. Type Frequency",  
 xlab="Frequency",  
 horiz=TRUE,  
 # Labels always: 0: parallel to axis, 1: horizontal,  
 # 2: perpendicular to axis, or 3: vertical.  
 las=1,  
 # Scale axis labels so all will be displayed.  
 cex.names=0.7  
)

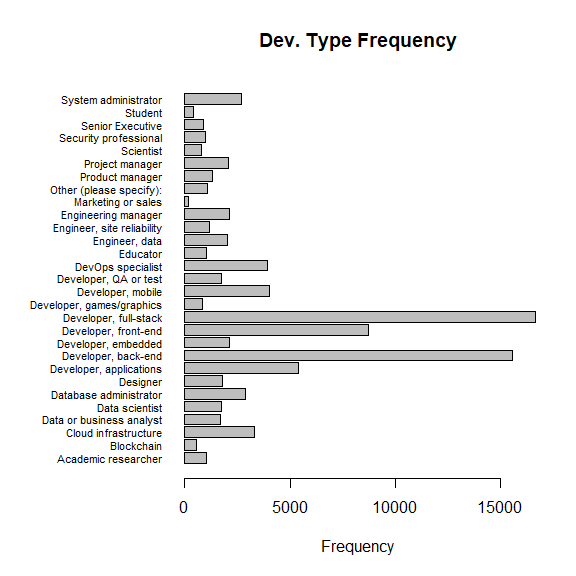


Figure 6-5: Dev. Type Frequency

Using ggscatter (Kassambara, 2023), a scatter plot produced a rather nebulous cloud toward the left-bottom of the chart. Its appearance is a little unusual because the x-axis is of an integral variable. Therefore, a regression line is added to help see the clustering of values (Figure 6-6).

ggscatter(  
 dat\_sliced, y="ConvertedCompYearly.n", x="YearsCodePro.n",  
 color="black", size=0.75,  
 conf.int=TRUE,  
 add.params=list(color="darkblue", fill="lightblue", size=0.5),  
 #color="group",  
 add="reg.line"  
) +  
 stat\_cor(method="pearson", label.x=30, label.y=-10000) +  
 labs(title="Yearly Compensation vs. Years Coding Professional\nScatter Plot with Linear Regression Line") +  
 scale\_y\_continuous(name="Compensation (US$)", labels=scales::dollar\_format())

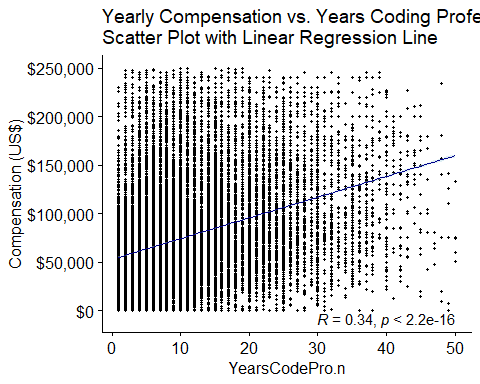


Figure 6-6: Yearly Compensation vs. Years Coding Professional Scatter Plot, fig.width=6

## 6.2 Normality of the Data

Using ggqqplot (Kassambara, 2023), a Q-Q (or Quantile-Quantile) plot is used to graphical interpret the normality of the Yearly Compensation distribution. Since the plotted values, it is apparent that the distribution is not normally distributed (Figure 6-7).

ggqqplot(dat\_sliced$ConvertedCompYearly.n, main="Yearly Compensation", xlab="Theoretical Quantiles")

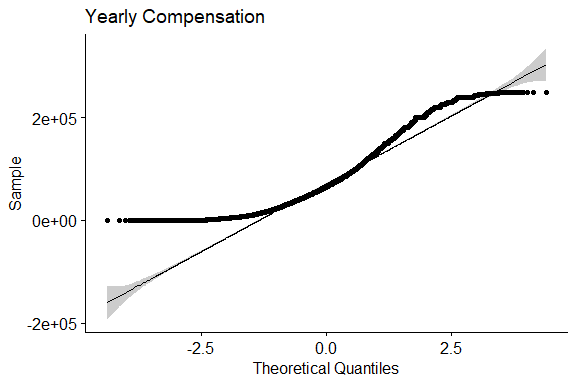


Figure 6-7: Yearly Compensation Q-Q Plot

The Q-Q plot reveals the data is not normally distributed, but a mathematical calculation is more precise. For this calculation, Shapiro-Wilk’s test (shapiro.test()) cannot be used because it supports sample size 3–5000 and the data contains nearly 90,000 samples. Additionally, the Kolmogorov-Smirnov test (ks.test()) does not support “ties” (in this case, two or more developers having the same annual compensation). Therefore, the Anderson-Darling test (ad.test()) was used because its main limitation is supporting a sample size of 7 or more (Table 6-1).

: ConvertedCompYearly.n is normally distributed, having p-value >= 0.05 (alpha).

: ConvertedCompYearly.n is not normally distributed, having p-value < 0.05.

ad.test(dat\_sliced$ConvertedCompYearly.n)

##   
## Anderson-Darling normality test  
##   
## data: dat\_sliced$ConvertedCompYearly.n  
## A = 1532.2, p-value < 2.2e-16

A p-value<0.001<0.05 rejects the null hypothesis, , indicating that the distribution of yearly compensation is not normally distributed.

## 6.3 Kurtosis and Skewness of the Data

The kurtosis is >3 indicating the distribution is leptokurtic, meaning there are more outliers than the normal distribution (Table 6-2).

kurtosis(dat\_sliced$ConvertedCompYearly.n)

## [1] 3.157151

The skewness is >1 indicating the distribution is highly right-skewed, meaning the curve’s “tail” tapers to the right (Table 6-3).

skewness(dat\_sliced$ConvertedCompYearly.n)

## [1] 0.851303

# 7 Results

## 7.1 Simple Linear Regressions

Compensation (ConvertedCompYearly.n) versus years coding professionally (YearsCodePro.n) linear regression model (Table 6-4).

compensationVersusYears\_lm <- lm(  
 # Dependent variable versus (~) independent variable(s).  
 ConvertedCompYearly.n ~ YearsCodePro.n,  
 data=dat\_sliced  
)  
summary(compensationVersusYears\_lm)

##   
## Call:  
## lm(formula = ConvertedCompYearly.n ~ YearsCodePro.n, data = dat\_sliced)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -157497 -37409 -10103 27998 190656   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 52697.8 273.7 192.5 <2e-16 \*\*\*  
## YearsCodePro.n 2138.9 19.6 109.2 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 50370 on 88835 degrees of freedom  
## Multiple R-squared: 0.1183, Adjusted R-squared: 0.1182   
## F-statistic: 1.191e+04 on 1 and 88835 DF, p-value: < 2.2e-16

Compensation (ConvertedCompYearly.n) versus gender (Gender.d) linear regression model (Table 6-5).

compensationVersusGender\_lm <- lm(  
 # Dependent variable versus (~) independent variable(s).  
 ConvertedCompYearly.n ~ Gender.d,  
 data=dat\_sliced  
)  
summary(compensationVersusGender\_lm)

##   
## Call:  
## lm(formula = ConvertedCompYearly.n ~ Gender.d, data = dat\_sliced)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -76281 -40670 -11794 30362 173638   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 76282.2 183.1 416.517 <2e-16 \*\*\*  
## Gender.dFemale -2288.5 988.2 -2.316 0.0206 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 53640 on 88835 degrees of freedom  
## Multiple R-squared: 6.036e-05, Adjusted R-squared: 4.91e-05   
## F-statistic: 5.362 on 1 and 88835 DF, p-value: 0.02058

Compensation (ConvertedCompYearly.n) versus organization size (OrgSize.log) linear regression model. Since OrgSize.log is the log of the original OrgSize, the result will be what is called a “level-log model” (because the model is vs. ) (Table 6-6).

compensationVersusOrgSize\_lm <- lm(  
 # Dependent variable versus (~) independent variable(s).  
 ConvertedCompYearly.n ~ OrgSize.log,  
 data=dat\_sliced  
)  
summary(compensationVersusOrgSize\_lm)

##   
## Call:  
## lm(formula = ConvertedCompYearly.n ~ OrgSize.log, data = dat\_sliced)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -91562 -40020 -11375 29588 189468   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 60401.14 363.27 166.27 <2e-16 \*\*\*  
## OrgSize.log 3022.78 60.63 49.86 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 52910 on 88835 degrees of freedom  
## Multiple R-squared: 0.02722, Adjusted R-squared: 0.02721   
## F-statistic: 2486 on 1 and 88835 DF, p-value: < 2.2e-16

Compensation (ConvertedCompYearly.n) versus developer type (DevType.f) linear regression model (Table 6-7).

compensationVersusDevType\_lm <- lm(  
 # Dependent variable versus (~) independent variable(s).  
 ConvertedCompYearly.n ~ DevType.f,  
 data=dat\_sliced  
)  
summary(compensationVersusDevType\_lm)

##   
## Call:  
## lm(formula = ConvertedCompYearly.n ~ DevType.f, data = dat\_sliced)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -107571 -40063 -10615 29671 212388   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 62295 1664 37.437 < 2e-16 \*\*\*  
## DevType.fBlockchain 18218 2758 6.606 3.98e-11 \*\*\*  
## DevType.fCloud infrastructure 31090 1899 16.370 < 2e-16 \*\*\*  
## DevType.fData or business analyst 11352 2098 5.410 6.33e-08 \*\*\*  
## DevType.fData scientist 18553 2094 8.859 < 2e-16 \*\*\*  
## DevType.fDatabase administrator 9848 1935 5.089 3.60e-07 \*\*\*  
## DevType.fDesigner 7487 2082 3.596 0.000323 \*\*\*  
## DevType.fDeveloper, applications 10601 1813 5.848 4.99e-09 \*\*\*  
## DevType.fDeveloper, back-end 13126 1717 7.645 2.11e-14 \*\*\*  
## DevType.fDeveloper, embedded 13236 2022 6.545 5.96e-11 \*\*\*  
## DevType.fDeveloper, front-end 5713 1757 3.251 0.001151 \*\*   
## DevType.fDeveloper, full-stack 11714 1714 6.836 8.18e-12 \*\*\*  
## DevType.fDeveloper, games/graphics 7923 2467 3.211 0.001323 \*\*   
## DevType.fDeveloper, mobile 2236 1860 1.202 0.229265   
## DevType.fDeveloper, QA or test 10450 2088 5.004 5.62e-07 \*\*\*  
## DevType.fDevOps specialist 23400 1865 12.546 < 2e-16 \*\*\*  
## DevType.fEducator 6543 2350 2.784 0.005372 \*\*   
## DevType.fEngineer, data 23816 2034 11.707 < 2e-16 \*\*\*  
## DevType.fEngineer, site reliability 36337 2274 15.983 < 2e-16 \*\*\*  
## DevType.fEngineering manager 45277 2022 22.395 < 2e-16 \*\*\*  
## DevType.fMarketing or sales 22615 4255 5.315 1.07e-07 \*\*\*  
## DevType.fOther (please specify): 25938 2306 11.245 < 2e-16 \*\*\*  
## DevType.fProduct manager 18507 2218 8.343 < 2e-16 \*\*\*  
## DevType.fProject manager 10836 2026 5.349 8.85e-08 \*\*\*  
## DevType.fScientist 15902 2502 6.357 2.07e-10 \*\*\*  
## DevType.fSecurity professional 29745 2366 12.573 < 2e-16 \*\*\*  
## DevType.fSenior Executive 41611 2447 17.005 < 2e-16 \*\*\*  
## DevType.fStudent -24814 3108 -7.984 1.44e-15 \*\*\*  
## DevType.fSystem administrator 9944 1950 5.100 3.40e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 52830 on 88808 degrees of freedom  
## Multiple R-squared: 0.03035, Adjusted R-squared: 0.03005   
## F-statistic: 99.28 on 28 and 88808 DF, p-value: < 2.2e-16

## 7.2 Multiple Linear Regression

The goal of the linear regression analysis was to test the null hypothesis, that all independent variables have a coefficient of zero:

:

:

Linear regression model:

Compensation (ConvertedCompYearly.n) versus years coding professionally (YearsCodePro.n), gender (Gender.d), organization size (OrgSize.log), and developer type (DevType.f) linear regression model (Table 6-8).

compensationVersusSignificants\_lm <- lm(  
 # Dependent variable versus (~) independent variables.  
 ConvertedCompYearly.n ~ OrgSize.log + Gender.d + YearsCodePro.n + DevType.f,  
 data=dat\_sliced  
)  
summary(compensationVersusSignificants\_lm)

##   
## Call:  
## lm(formula = ConvertedCompYearly.n ~ OrgSize.log + Gender.d +   
## YearsCodePro.n + DevType.f, data = dat\_sliced)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -161014 -35161 -9444 26220 203834   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 20904.33 1590.01 13.147 < 2e-16 \*\*\*  
## OrgSize.log 3180.40 56.81 55.983 < 2e-16 \*\*\*  
## Gender.dFemale 2672.13 901.75 2.963 0.003045 \*\*   
## YearsCodePro.n 2152.26 19.29 111.558 < 2e-16 \*\*\*  
## DevType.fBlockchain 27675.78 2551.14 10.848 < 2e-16 \*\*\*  
## DevType.fCloud infrastructure 30668.58 1755.27 17.472 < 2e-16 \*\*\*  
## DevType.fData or business analyst 8851.39 1939.41 4.564 5.03e-06 \*\*\*  
## DevType.fData scientist 20875.40 1935.23 10.787 < 2e-16 \*\*\*  
## DevType.fDatabase administrator 8789.74 1791.14 4.907 9.25e-07 \*\*\*  
## DevType.fDesigner 6793.18 1926.76 3.526 0.000423 \*\*\*  
## DevType.fDeveloper, applications 6762.82 1676.17 4.035 5.47e-05 \*\*\*  
## DevType.fDeveloper, back-end 14671.90 1586.82 9.246 < 2e-16 \*\*\*  
## DevType.fDeveloper, embedded 11111.17 1869.31 5.944 2.79e-09 \*\*\*  
## DevType.fDeveloper, front-end 10258.12 1624.89 6.313 2.75e-10 \*\*\*  
## DevType.fDeveloper, full-stack 15173.57 1584.23 9.578 < 2e-16 \*\*\*  
## DevType.fDeveloper, games/graphics 10911.81 2281.62 4.782 1.73e-06 \*\*\*  
## DevType.fDeveloper, mobile 7079.59 1720.87 4.114 3.89e-05 \*\*\*  
## DevType.fDeveloper, QA or test 11197.75 1929.80 5.803 6.55e-09 \*\*\*  
## DevType.fDevOps specialist 22509.24 1723.84 13.058 < 2e-16 \*\*\*  
## DevType.fEducator 4652.99 2172.65 2.142 0.032227 \*   
## DevType.fEngineer, data 23729.27 1879.78 12.623 < 2e-16 \*\*\*  
## DevType.fEngineer, site reliability 35038.91 2101.00 16.677 < 2e-16 \*\*\*  
## DevType.fEngineering manager 39729.62 1869.77 21.248 < 2e-16 \*\*\*  
## DevType.fMarketing or sales 23471.06 3933.61 5.967 2.43e-09 \*\*\*  
## DevType.fOther (please specify): 20987.76 2131.69 9.846 < 2e-16 \*\*\*  
## DevType.fProduct manager 17096.88 2054.05 8.323 < 2e-16 \*\*\*  
## DevType.fProject manager 10688.37 1875.55 5.699 1.21e-08 \*\*\*  
## DevType.fScientist 13462.38 2311.55 5.824 5.77e-09 \*\*\*  
## DevType.fSecurity professional 25796.25 2187.18 11.794 < 2e-16 \*\*\*  
## DevType.fSenior Executive 38458.63 2268.84 16.951 < 2e-16 \*\*\*  
## DevType.fStudent -9413.98 2874.69 -3.275 0.001058 \*\*   
## DevType.fSystem administrator 9939.65 1803.87 5.510 3.59e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 48820 on 88805 degrees of freedom  
## Multiple R-squared: 0.1721, Adjusted R-squared: 0.1719   
## F-statistic: 595.7 on 31 and 88805 DF, p-value: < 2.2e-16

## 7.3 Interpretation

With a p-value<0.001<0.05 (2.2\*10^-16) and an F-stat.=595.7 (d.f.=88,805), the model is statistically significant and can be used to reliably predict Yearly Compensation. Therefore, the null hypothesis (: ) which assumes there is no linear relationship is rejected in favor of .

A R=17.21% indicates there is a weak, positive linear correlation between the observed values and the predicted values. An =17.19% indicates that only 17.19% of the variance in ConvertedCompYearly.n () can be explained by the variances in YearsCodePro.n (), Gender.d (), OrgSize.log (), and DevType.f ().

**The coefficient of OrgSize.log is =3,180.40:** The results indicate that a percentage increase in Org. Size results in a US$31.80 () increase in Yearly Compensation, holding all other variables constant. This is significantly different from 0 since p-value<0.001<0.05 (; ).

**The coefficient of Gender.d is =2,672.13:** Since Gender is coded 0/1 (0=Other; 1=Female), the interpretation is that for , a US$2,672.13 unit increase in Yearly Compensation is predicted (holding all other variables constant). This is significantly different from 0 since p-value=0.003<0.05.

**The coefficient of YearsCodePro.n is =2,152.26:** For every 1 unit (years) increase in YearsCodePro.n, a US$2,152.26 unit increase in Yearly Compensation is predicted (holding all other variables constant). This is significantly different from 0 since p-value<0.001<0.05.

All the coefficients of DevType.f are significantly different from 0 since every (all 28 Dev. Type factors) have p-value<0.05.

Linear regression equation:

## 7.4 Example Prediction

For example, the average IT professional who works for a company with ~300 employees, is female, has 8 years experience (coding professionally), and works in Marketing could expect a yearly compensation of . Furthermore, since the survey is from IT professionals all over the world, it is a fairly reasonable assumption that the compensation could be multiplied by the cost of living index (or COLI) to adjust expectations. If the IT professional lived in an area with a COLI of 74.1 (that of Murray State University, Murray, Kentucky), a reasonable geographic expectation might be more like .

# 8 Implications

With the discussion of “equal pay for equal work” (read “equitable pay for similar work”) evolving into countless litigation, it would seem that the work environment is decidedly sexist and non-males are compensated less for the same work. For example, the U.S. Women’s National Soccer Team has brought suit for pay discrimination every year from 2019 to 2023. The media coverage of this gives the impression that they have won, but the reality is much more complicated (Dure, 2022).

It seems that giving the public the impression that women are not equitably compensated has resulted in the opposite result, perhaps because compensation was equitable before the response in the workplace. Statistical results like this should be referenced when developing workplace policies and practices, instead of referencing a news anchor’s opinion or someone’s feelings expressed during a press conference.

Another possibility is that there are more Female IT professionals in more urban areas than rural areas. These urban areas would most likely have a higher COLI than rural areas thereby resulting in higher yearly compensation. However, the source dataset does not include geographical information beyond country. Further analysis could be performed with a variable indicating geography, such as FIPS (U.S. specific), postal code with country, or a global solution like “Geohash” (Morton, 1966).

# 9 Conclusion

This study examined the association of yearly compensation in IT positions against gender and three other variables: organization size, years coding professionally, and developer type. Using multiple linear regression analysis, it was concluded that not only does gender make a statistically significant difference in yearly compensation, but female IT professionals receive more yearly compensation than non-females, holding all other variables constant.

# 10 References

Bache, S., Wickham, H. (2022). *magrittr: A Forward-Pipe Operator for R*. <https://CRAN.R-project.org/package=magrittr>

Bajarin, Tim (April 14, 2020). *Four Waves of Personal Computing and the PC’s New Role as an Essential Tool for Survival*. Retrieved April 28, 2024 from <https://www.forbes.com/sites/timbajarin/2020/04/14/four-waves-of-personal-computing-and-the-pcs-new-role-as-an-essential-tool-for-survival/>

Ballenstedt, Brittany (April 29, 2010). *Tech Jobs Declined 4 Percent in 2009*. Retrieved April 27, 2024 from <https://www.nextgov.com/digital-government/2010/04/tech-jobs-declined-4-percent-in-2009/53257/>

Carlow University (December 29, 2022). *How the Gender Pay Gap Impacts Female Nurses*. Retrieved April 28, 2024 from <https://blog.carlow.edu/2022/12/29/how-the-gender-pay-gap-impacts-female-nurses/>

Coursera Staff (March 13, 2024). *IT Salary Overview: How Much Can You Make?* Retrieved April 28, 2024 from <https://www.coursera.org/articles/it-salaries-roles-location-and-experience>

Davis, Jessica (May 25, 2022). *Salary Survey: IT Gender Pay Gap Keeps Growing*. Retrieved April 28, 2024 from <https://www.informationweek.com/it-leadership/salary-survey-it-gender-pay-gap-keeps-growing>

Dure, Beau (December 16, 2022). *Did the US women’s soccer team win their ‘equal pay’ case? It’s complicated*. Retrieved May 8, 2024 from <https://www.theguardian.com/football/2022/dec/16/uswnt-equal-pay-settlement-case>

Gross, J., Ligges, U. (2015). *nortest: Tests for Normality*. <https://CRAN.R-project.org/package=nortest>

Haan, Katherine (March 1, 2024). *Gender Pay Gap Statistics in 2024*. Retrieved April 27, 2024 from <https://www.forbes.com/advisor/business/gender-pay-gap-statistics/>

Hlavac, Marek (2022). *stargazer: Well-Formatted Regression and Summary Statistics Tables*. <https://CRAN.R-project.org/package=stargazer>

Inspirisys Solutions (March 5, 2024). *Tech layoffs in 2023: Causes, Consequences and Affected Companies*. Retrieved April 27, 2024 from <https://www.linkedin.com/pulse/tech-layoffs-2023-causes-consequences-affected-companies-inspirisys-bbc6c>

Kassambara, A. (2023). *ggpubr: Based Publication Ready Plots*. <https://CRAN.R-project.org/package=ggpubr>

Komsta, L., Novomestky, F. (2022). *moments: Moments, Cumulants, Skewness, Kurtosis and Related Tests*. <https://CRAN.R-project.org/package=moments>

Morton, G. M. (March 1, 1966). *A Computer Oriented Geodetic Data Base and a New Technique in File Sequencing*. International Business Machines Co. Ltd., Canada.

Nguyen, Hieu (December 10, 2022). *Is Outsourcing Good or Bad? The Economic Effects of Outsourcing*. Retrieved April 27, 2024 from <https://www.orientsoftware.com/blog/is-outsourcing-good-or-bad/>

Nichols, A. & Zimmerman, S. (May 2008). *Measuring Trends in Income Variability*. Retrieved April 20, 2024 from <https://www.urban.org/sites/default/files/publication/31811/411688-measuring-trends-in-income-variability.pdf>

Shim, Richard (November 19, 2003). *A half-million tech jobs lost in 2002, study says*. Retrieved April 27, 2024 from <https://www.cnet.com/culture/a-half-million-tech-jobs-lost-in-2002-study-says/>

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. <https://CRAN.R-project.org/package=ggplot2>

Wickham, H., Vaughan, D., Girlich, M. (2024). *tidyr: Tidy Messy Data*. <https://CRAN.R-project.org/package=tidyr>

Wooldridge, Jeffrey (2008). *Introductory Economics: A Modern Approach* (4th ed.). Cengage Learning, Inc.