Analysis of Annual Compensation in Information Technology

CIS663-01-SP24 Term Project Report

Mark M. Young (M00348726)

May 09, 2024

Table of Contents

# 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.

# 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).

Lorem ipsum

# 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

Lorem ipsum

# 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. Capture logarithmic values in case it is a multimodal distribution that needs to be split.  
dat\_selected$ConvertedCompYearly.n <- as.numeric(as.character(dat\_selected$ConvertedCompYearly))  
dat\_selected$ConvertedCompYearly.log <- log(dat\_selected$ConvertedCompYearly.n)

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

## Warning: NAs introduced by coercion

### 5.4.2 Categorical Variables

Lorem ipsum

# Copy variable.  
dat\_selected$Gender.f <- dat\_selected$Gender  
# Replace all values that are not "Man", "Woman", or NA with "Other".  
dat\_selected$Gender.f[!dat\_selected$Gender %in% c("Man", "Woman", NA)] <- "Other"  
# Assign dummy values.  
dat\_selected$Gender.f[dat\_selected$Gender.f == "Man"] <- 0  
dat\_selected$Gender.f[dat\_selected$Gender.f == "Woman"] <- 1  
dat\_selected$Gender.f[dat\_selected$Gender.f == "Other"] <- 2  
# Explicitly convert.  
dat\_selected$Gender.f <- as.integer(dat\_selected$Gender.f)  
dat\_selected$Gender.f <- factor(  
 dat\_selected$Gender.f,  
 levels=c(0, 1, 2),  
 labels=c("Male", "Female", "Other")  
)

Lorem ipsum

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

### 5.4.3 Ordinal Variables

Lorem ipsum

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

Replacement of character values with integer values omitted for brevity.

dat\_selected$OrgSize.o <- ordered(  
 dat\_selected$OrgSize.o,  
 exclude=NULL,  
 levels=c(  
 0, 1, 5, 15, 60,  
 300, 750, 3000, 7500, 30000  
 ),  
 labels=c(  
 "Unknown", "1", "2--9", "10--19", "20--99",  
 "100--499", "500--999", "1,000--4,999", "5,000--9,999", "10,000+"  
 )  
)

## 5.5 Value Filtering

Lorem ipsum

# Remove observations without values in the variables in interest.  
dat\_sliced <- dat\_selected[  
 !is.na(dat\_selected$ConvertedCompYearly.n)  
 & !is.na(dat\_selected$YearsCodePro.n)  
 & !is.na(dat\_selected$Gender.f)  
 & !is.na(dat\_selected$OrgSize.o)  
 & !is.na(dat\_selected$DevType.f)  
 # Be sure to specify an comma at the end to avoid "undefined columns selected" error.  
 ,  
]  
  
# Remove outliers.  
dat\_sliced <- dat\_sliced[dat\_sliced$ConvertedCompYearly.n < 250000,]

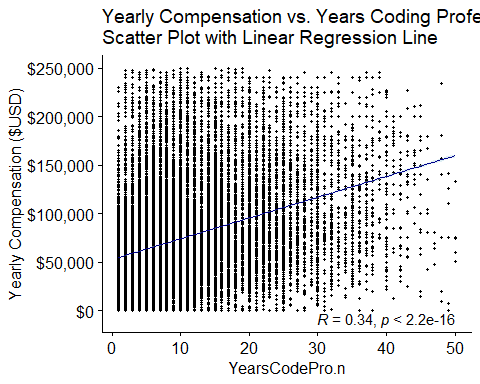
# 6 Methodology

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

Lorem ipsum

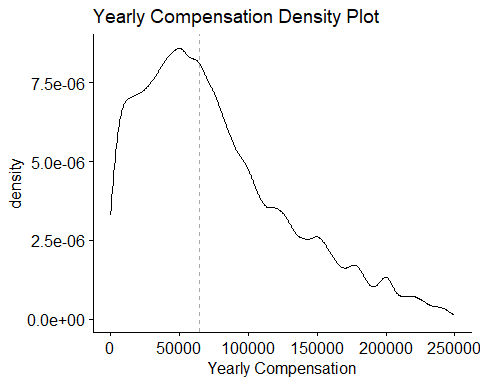
ggscatter(  
 dat\_sliced, y="ConvertedCompYearly.n", x="YearsCodePro.n",  
 color="black", size=0.75,  
 conf.int=TRUE,  
 add.params=list(color="darkblue", fill="lightblue", size=0.5),  
 #color="group",  
 add="reg.line"  
) +  
 stat\_cor(method="pearson", label.x=30, label.y=-10000) +  
 labs(title="Yearly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



Figure

Lorem ipsum

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



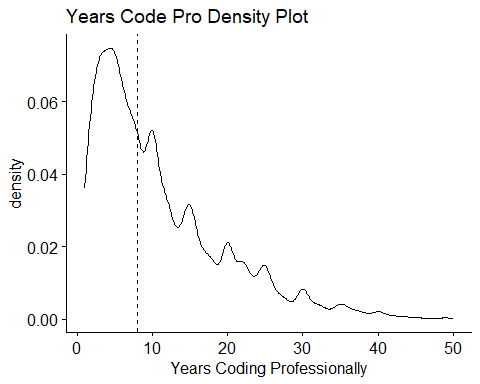
Figure

Lorem ipsum

ggdensity(dat\_sliced$YearsCodePro.n,  
 main="Years Code Pro Density Plot",  
 xlab="Years Coding Professionally",  
 add="median"  
) #+

## Warning: `geom\_vline()`: Ignoring `mapping` because `xintercept` was provided.

## Warning: `geom\_vline()`: Ignoring `data` because `xintercept` was provided.

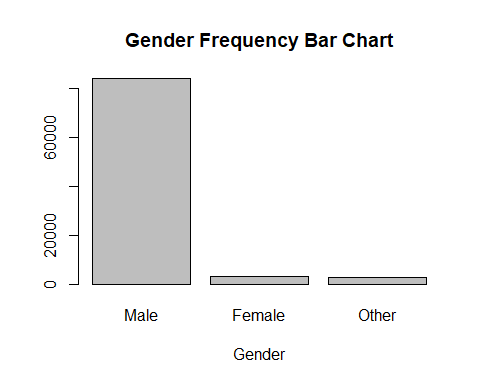


Figure

#geom\_vline(aes(xintercept=median(dat\_sliced$YearsCodePro.n)),  
# colour="darkgray", linetype="dashed"  
#)

Lorem ipsum

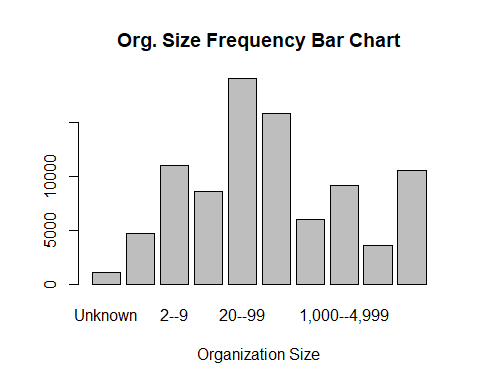
f.gender <- barplot(  
 table(dat\_sliced$Gender.f),  
 main="Gender Frequency Bar Chart",  
 xlab="Gender"  
)



Figure

Lorem ipsum

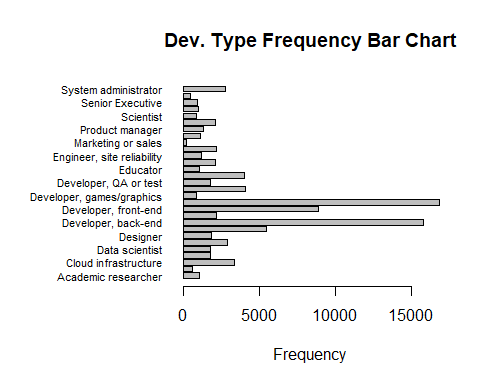
f.orgsize <- barplot(  
 table(dat\_sliced$OrgSize.o),  
 main="Org. Size Frequency Bar Chart",  
 xlab="Organization Size"  
)



Figure

Lorem ipsum

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 Bar Chart",  
 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  
)



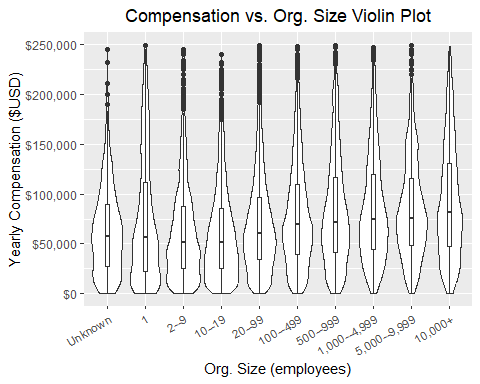
Figure

Lorem ipsum

#TODO ConvertedCompYearly vs. YearsCodePro scatter plot

OrgSize.o is not statistically significant, but an upward trend is apparent. Lorem ipsum

dat\_comp\_vs\_org\_df <- dat\_sliced[c("ConvertedCompYearly.n", "OrgSize.o")]  
violin <- ggplot(dat\_comp\_vs\_org\_df, aes(y=ConvertedCompYearly.n, x=OrgSize.o))  
violin +  
 geom\_violin() +  
 geom\_boxplot(width=0.1, aes(middle=mean(ConvertedCompYearly.n))) +  
 labs(title="Compensation vs. Org. Size Violin Plot", x="Org. Size (employees)") +  
 scale\_y\_continuous(name="Yearly Compensation ($USD)", labels=scales::dollar\_format()) +  
 theme(axis.text.x=element\_text(angle=30, hjust=1), plot.title=element\_text(hjust=0.5))

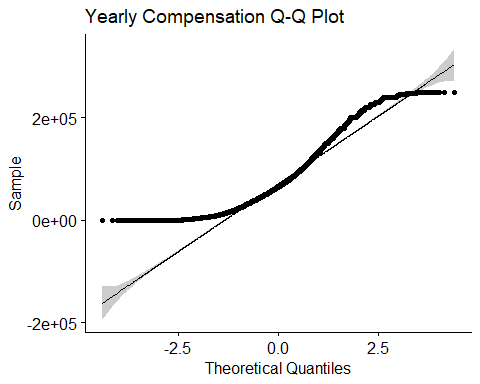


Figure

## 6.2 Normality of the Data

Lorem ipsum

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



Figure

The Q-Q (quantile-quantile) plot reveals the data is not normally distributed because of the deviation from the straight line. A mathematical calculation is still 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). Therefor, the Anderson-Darling test (ad.test()) was used because its main limitation is supporting a sample size of 7 or more.

: 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 = 1548, p-value < 2.2e-16

## 6.3 Kurtosis and Skewness of the Data

kurtosis(dat\_sliced$ConvertedCompYearly.n)

## [1] 3.166628

The kurtosis is >3 indicating the distribution is leptokurtic, meaning there are more outliers than the normal distribution.

skewness(dat\_sliced$ConvertedCompYearly.n)

## [1] 0.8533877

The skewness is between 0.5 and 1.0 indicating the distribution is moderately right-skewed, meaning the curve’s “tail” tapers to the right.

# 7 Results

## 7.1 Simple Linear Regressions

Compensation (ConvertedCompYearly.n) versus years coding professionally (YearsCodePro.n) linear regression model.

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   
## -157523 -37317 -10030 27742 190799   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 52547.80 271.41 193.6 <2e-16 \*\*\*  
## YearsCodePro.n 2142.51 19.46 110.1 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 50270 on 89938 degrees of freedom  
## Multiple R-squared: 0.1188, Adjusted R-squared: 0.1188   
## F-statistic: 1.212e+04 on 1 and 89938 DF, p-value: < 2.2e-16

Compensation (ConvertedCompYearly.n) versus gender (Gender.f) linear regression model.

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

##   
## Call:  
## lm(formula = ConvertedCompYearly.n ~ Gender.f, data = dat\_sliced)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -81084 -40774 -11818 30678 173903   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 75966.5 184.6 411.571 < 2e-16 \*\*\*  
## Gender.fFemale -2027.0 979.7 -2.069 0.0386 \*   
## Gender.fOther 5120.5 1050.1 4.876 1.08e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 53550 on 89937 degrees of freedom  
## Multiple R-squared: 0.0003196, Adjusted R-squared: 0.0002974   
## F-statistic: 14.38 on 2 and 89937 DF, p-value: 5.708e-07

Compensation (ConvertedCompYearly.n) versus organization size (OrgSize.o) linear regression model. Note, L, Q, C, ^4, ^5, etc. indicate Linear, Quadratic, Cubic, fourth-degree, fifth-degree, etc.

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

##   
## Call:  
## lm(formula = ConvertedCompYearly.n ~ OrgSize.o, data = dat\_sliced)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -92527 -39729 -11063 29757 181856   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 75909.4 240.4 315.725 < 2e-16 \*\*\*  
## OrgSize.o.L 28733.5 972.9 29.534 < 2e-16 \*\*\*  
## OrgSize.o.Q 6302.8 945.9 6.664 2.69e-11 \*\*\*  
## OrgSize.o.C -4977.1 879.4 -5.659 1.52e-08 \*\*\*  
## OrgSize.o^4 -1546.6 798.7 -1.937 0.0528 .   
## OrgSize.o^5 11057.0 777.4 14.223 < 2e-16 \*\*\*  
## OrgSize.o^6 -5309.8 665.4 -7.979 1.49e-15 \*\*\*  
## OrgSize.o^7 1796.6 575.7 3.121 0.0018 \*\*   
## OrgSize.o^8 1091.7 580.4 1.881 0.0600 .   
## OrgSize.o^9 -245.7 484.4 -0.507 0.6119   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 52660 on 89930 degrees of freedom  
## Multiple R-squared: 0.03322, Adjusted R-squared: 0.03313   
## F-statistic: 343.4 on 9 and 89930 DF, p-value: < 2.2e-16

Compensation (ConvertedCompYearly.n) versus developer type (DevType.f) linear regression model.

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   
## -107431 -39971 -10551 29565 212725   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 61541 1619 38.006 < 2e-16 \*\*\*  
## DevType.fBlockchain 18412 2715 6.782 1.19e-11 \*\*\*  
## DevType.fCloud infrastructure 31701 1858 17.066 < 2e-16 \*\*\*  
## DevType.fData or business analyst 11901 2056 5.787 7.18e-09 \*\*\*  
## DevType.fData scientist 19028 2051 9.276 < 2e-16 \*\*\*  
## DevType.fDatabase administrator 10484 1894 5.535 3.11e-08 \*\*\*  
## DevType.fDesigner 8415 2041 4.122 3.75e-05 \*\*\*  
## DevType.fDeveloper, applications 11262 1769 6.365 1.96e-10 \*\*\*  
## DevType.fDeveloper, back-end 13770 1673 8.232 < 2e-16 \*\*\*  
## DevType.fDeveloper, embedded 13749 1980 6.946 3.79e-12 \*\*\*  
## DevType.fDeveloper, front-end 6369 1713 3.717 0.000202 \*\*\*  
## DevType.fDeveloper, full-stack 12394 1669 7.424 1.14e-13 \*\*\*  
## DevType.fDeveloper, games/graphics 8667 2427 3.571 0.000356 \*\*\*  
## DevType.fDeveloper, mobile 2856 1818 1.571 0.116113   
## DevType.fDeveloper, QA or test 11284 2047 5.512 3.55e-08 \*\*\*  
## DevType.fDevOps specialist 24000 1823 13.166 < 2e-16 \*\*\*  
## DevType.fEducator 7006 2304 3.041 0.002356 \*\*   
## DevType.fEngineer, data 24206 1993 12.148 < 2e-16 \*\*\*  
## DevType.fEngineer, site reliability 36763 2233 16.465 < 2e-16 \*\*\*  
## DevType.fEngineering manager 45891 1983 23.144 < 2e-16 \*\*\*  
## DevType.fMarketing or sales 23193 4222 5.494 3.95e-08 \*\*\*  
## DevType.fOther (please specify): 26350 2264 11.641 < 2e-16 \*\*\*  
## DevType.fProduct manager 19229 2182 8.812 < 2e-16 \*\*\*  
## DevType.fProject manager 11513 1986 5.797 6.79e-09 \*\*\*  
## DevType.fScientist 15929 2445 6.515 7.31e-11 \*\*\*  
## DevType.fSecurity professional 30476 2330 13.079 < 2e-16 \*\*\*  
## DevType.fSenior Executive 42306 2414 17.526 < 2e-16 \*\*\*  
## DevType.fStudent -24397 3033 -8.044 8.80e-16 \*\*\*  
## DevType.fSystem administrator 10712 1908 5.614 1.98e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 52740 on 89911 degrees of freedom  
## Multiple R-squared: 0.03033, Adjusted R-squared: 0.03003   
## F-statistic: 100.4 on 28 and 89911 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.

:

:

Compensation (ConvertedCompYearly.n) versus years coding professionally (YearsCodePro.n), gender (Gender.f), organization size (OrgSize.o), and developer type (DevType.f) linear regression model.

Linear regression model:

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

##   
## Call:  
## lm(formula = ConvertedCompYearly.n ~ YearsCodePro.n + Gender.f +   
## OrgSize.o + DevType.f, data = dat\_sliced)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -158940 -34875 -9257 26225 203318   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 37504.58 1511.26 24.817 < 2e-16 \*\*\*  
## YearsCodePro.n 2136.98 19.31 110.679 < 2e-16 \*\*\*  
## Gender.fFemale 2877.00 892.72 3.223 0.001270 \*\*   
## Gender.fOther 5054.63 954.86 5.294 1.20e-07 \*\*\*  
## OrgSize.o.L 28483.77 901.79 31.586 < 2e-16 \*\*\*  
## OrgSize.o.Q 6048.03 876.21 6.902 5.15e-12 \*\*\*  
## OrgSize.o.C -8070.52 816.11 -9.889 < 2e-16 \*\*\*  
## OrgSize.o^4 5482.20 741.80 7.390 1.48e-13 \*\*\*  
## OrgSize.o^5 5324.02 720.69 7.387 1.51e-13 \*\*\*  
## OrgSize.o^6 -2363.37 615.83 -3.838 0.000124 \*\*\*  
## OrgSize.o^7 667.05 532.20 1.253 0.210069   
## OrgSize.o^8 448.75 536.38 0.837 0.402804   
## OrgSize.o^9 -685.12 447.61 -1.531 0.125868   
## DevType.fBlockchain 26879.39 2508.25 10.716 < 2e-16 \*\*\*  
## DevType.fCloud infrastructure 30343.22 1716.07 17.682 < 2e-16 \*\*\*  
## DevType.fData or business analyst 8492.69 1898.84 4.473 7.74e-06 \*\*\*  
## DevType.fData scientist 20544.55 1893.54 10.850 < 2e-16 \*\*\*  
## DevType.fDatabase administrator 8650.64 1751.96 4.938 7.92e-07 \*\*\*  
## DevType.fDesigner 6648.03 1887.49 3.522 0.000428 \*\*\*  
## DevType.fDeveloper, applications 6418.85 1635.12 3.926 8.66e-05 \*\*\*  
## DevType.fDeveloper, back-end 14278.40 1545.22 9.240 < 2e-16 \*\*\*  
## DevType.fDeveloper, embedded 10686.07 1828.38 5.845 5.10e-09 \*\*\*  
## DevType.fDeveloper, front-end 9924.32 1583.27 6.268 3.67e-10 \*\*\*  
## DevType.fDeveloper, full-stack 14836.39 1542.67 9.617 < 2e-16 \*\*\*  
## DevType.fDeveloper, games/graphics 10511.20 2241.96 4.688 2.76e-06 \*\*\*  
## DevType.fDeveloper, mobile 6749.46 1680.48 4.016 5.91e-05 \*\*\*  
## DevType.fDeveloper, QA or test 11070.15 1890.16 5.857 4.74e-09 \*\*\*  
## DevType.fDevOps specialist 22152.23 1683.83 13.156 < 2e-16 \*\*\*  
## DevType.fEducator 3972.77 2126.71 1.868 0.061761 .   
## DevType.fEngineer, data 23150.92 1839.76 12.584 < 2e-16 \*\*\*  
## DevType.fEngineer, site reliability 34304.25 2061.40 16.641 < 2e-16 \*\*\*  
## DevType.fEngineering manager 39380.65 1833.55 21.478 < 2e-16 \*\*\*  
## DevType.fMarketing or sales 22681.92 3897.74 5.819 5.93e-09 \*\*\*  
## DevType.fOther (please specify): 20227.53 2089.52 9.680 < 2e-16 \*\*\*  
## DevType.fProduct manager 17096.48 2019.40 8.466 < 2e-16 \*\*\*  
## DevType.fProject manager 10665.28 1838.18 5.802 6.57e-09 \*\*\*  
## DevType.fScientist 13209.81 2255.49 5.857 4.74e-09 \*\*\*  
## DevType.fSecurity professional 25384.60 2152.15 11.795 < 2e-16 \*\*\*  
## DevType.fSenior Executive 39125.35 2237.85 17.483 < 2e-16 \*\*\*  
## DevType.fStudent -9769.17 2800.67 -3.488 0.000487 \*\*\*  
## DevType.fSystem administrator 9860.41 1764.08 5.590 2.28e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 48650 on 89899 degrees of freedom  
## Multiple R-squared: 0.1751, Adjusted R-squared: 0.1747   
## F-statistic: 477.1 on 40 and 89899 DF, p-value: < 2.2e-16

With a p-value<0.001 (2.2\*10^-16), an F-stat.=477.1, and an Adj-R2=17.47%, the model is statistically significant and only 17.47% of the variance in ConvertedCompYearly.n () can be explained by the variances in YearsCodePro.n (), Gender.f (), OrgSize.o (), and DevType.f ().

Linear regression equation:

## 7.3 Example

For example, the average IT professional with 8 years experience (coding professionally), is Female, in an organization with ~300 employees, 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 to adust expectations. If the IT professional lived in an area with a cost of living index of 74.1 (that of Murray State University), 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.

# 9 Conclusion

This study examined the association of yearly compensation in IT positions against gender and three other variables: years coding professorially, organization size, 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 females make more than men and those who identify as a gender other than male or female make even more than females.

# 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/> # #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>

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>

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>

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>