An Analysis of Total Compensation of People and its Correlation to Age they First Started to Code

*B\_Group 30*

*Rajat Shrestha, Pranav Khatri, Sanjog Chaudhary,*

*Darpan Pokhrel, Aashish Upadhyaya*

Friday, 14 Jan 2022

**Abstract:**

Context: Tech is the highest paid profession all over the world. Every day younger developers get into the market who are more exposed to the digital elements, this research is to find if starting to code earlier will affect their earning potential.

Question: This study asks if there is any correlation between the age an individual first started to code and their total yearly income?

Method: We analyzed the reported total annual compensation of people (in USD) of thousands of respondents of the stack overflow developer survey 2020.

Results: We expected to see the relation of the age of the developer when they started to code to influence their net income but after analyzing the dataset, we found no strong relationship between these factors.

Conclusion: The results suggest that there is no correlation between the age an individual first started to code and their total annual compensation.

# Introduction

The cash rewards paid to the employees in exchange for their services is called compensation (Cambridge University Press, 1995). It is common to measure compensation on basis of annual total income (MaRS, 2022).

Coding is an act of writing computer instructions (Cambridge University Press, 1995).

This study aims to understand if there is any relation to the age they started to code for their compensation. Specifically, we ask the following research question: Is there a correlation between total compensation and age first started to code?

The null hypothesis is:

**H0**: There is no correlation between total compensation and age first started to code.

The alternative hypothesis is:

**Halt**: There is a correlation between total compensation and age first started to code.

The data set we used is “Stack Overflow Developer Survey 2020”. It is the largest public survey participated by developers all around the world (Stack Overflow, 2022). In 2020, 65000 participants took part in this survey and we chose this year’s data as there was a column that gave the data about the annual total compensation of each participant in USD and a column that indicated at what age they first started to code.

The URL of the dataset is: “<https://data.world/technology/stack-overflow-developer-survey/workspace/file?filename=2020-survey_results_public.csv>”

Results: The results are discussed more in detail after the Visualization and Analysis phase where the plots and results point out to have no statistical significance between these variables.

In the rest of this paper, we present a visual view of the data, followed by statistical analysis, and end the report with a discussion of the implications of the results.

A picture containing qr code

Description automatically generated

Figure : The official stack overflow Developer survey overview (Stack Overflow, 2022).

# Visualization

The dataset used for this study comprises 64,461 rows (entry of each respondent) and 62 columns (different data points of each respondent). The age of the developers since they started to code is limited from 0 to 90. The maximum compensation threshold is set to 200000 USD.

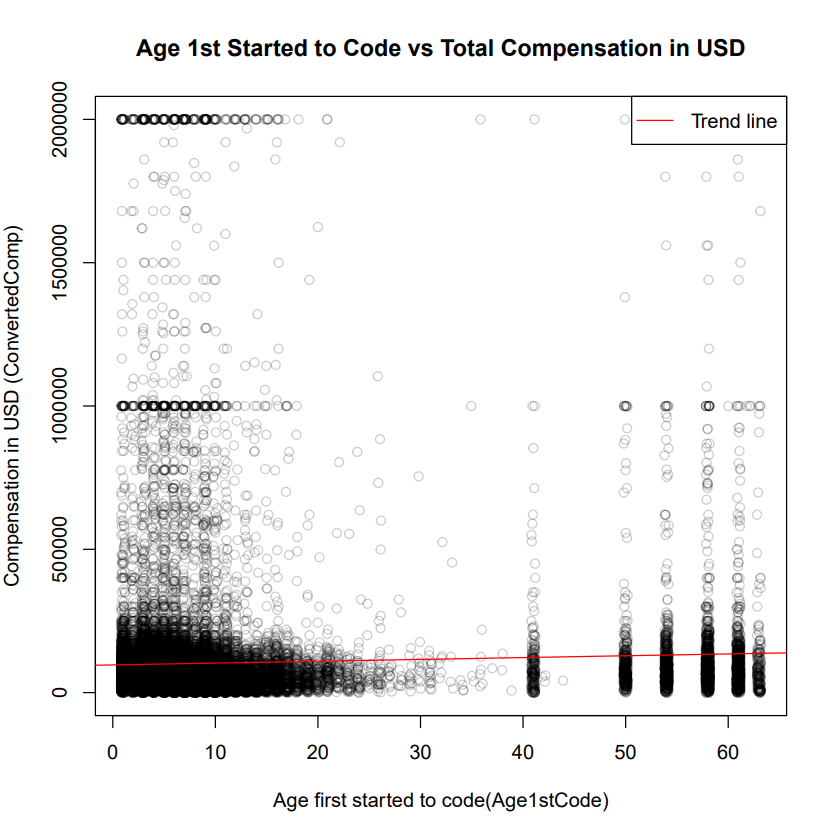


Figure : Age First Started to code vs Total compensation graph with trend line

Figure 2 plots all the entries by everyone as a single point in a graph with horizontal distance from the left suggesting the age the individual first started to code and the vertical distance from the bottom depicting their total compensation. The “red” line highlights the trend to describe more clearly what is happening in the graph. It appears that the trend line is higher at the right and gradually declines at the left suggesting that the people who started to code early have relatively less total compensation than those who started coding later in their life. It is quite unusual to see a large percentage of people starting to code after 40 years and earning more than the younger ones. The diagram also shows that there are a certain group of people who started to learn coding a lot earlier (roughly from age 3-13) who are earning a lot compared to the other demographic.

Graphical user interface

Description automatically generated with low confidence

Figure : Frequency Distribution of the Total compensation

Figure 3 depicts a histogram showing the frequency distribution of total compensation earned by all the Respondents. This histogram shows that the total compensation is skewed toward the lower end, this is unusual as normal distribution would follow a rather distinct “bell curve” (Herrnstein & Murray, 1994). We can see that most of the respondents have compensation on the lower end. There is some increase in frequencies at the 1-million and 2-million dollar mark but most of the developers reported total compensation of fewer than 500,000 dollars per annum.

# Analysis

# 

Figure : Output of the analysis.R file

Figure 3 clearly shows that the total compensation is not normally distributed, so we used the spearman function provided by R statistical to test the correlation (Lund Research Ltd, 2018) (R Core Team, 2021). This test performs a spearman test which evaluates the monotonic relationship between two continuous or ordinal variables to figure out if there is any relation between these two discrete variables (Spearman, 1904). Comparing the change of total compensation concerning their age.

From Figure 4 we can see that the s-value is huge (7.544e+12), and the p-value is extremely small (2.2e-16). The big S value suggests the variance in the data itself while the p-value expresses the degree of change on both variables and conveys a relationship between the variables (Andrade, 2019). But in this case, since the p-value is tiny (2.2e-16) this suggests that there is no statistical significance of these two variables. Finally, we get the rho value which is negative and smaller than 0.1 in magnitude which means that there is no or negligible correlation between these values (Spearman, 1904).

Thus, the null hypothesis cannot be rejected in favour of the alternative hypothesis as the observed P-value and rho estimates are smaller than the acceptable range.

# Conclusions

From the plotting of the age first code vs total compensation, we can see a lot of respondents scattered all around the chart with a high degree of variance in between them. The trend line seems almost parallel to the horizontal axis which was suggesting no significant correlation between these variables. Along which we can see three distinct groups of people in this result. The most common one is the average developer who starts coding from 7 to 15 who earns the average salary, a significant number of people who started to write code over 40 years who have early coders who also earn the same average salary, and some people who started to code early and earn a lot of money.

Due to the huge boost in technology, there is a big boom in the technology sector which is now demanding a lot of skilled IT professionals, increasing their demand in the big tech hubs and even in rising local businesses causing their compensation to be higher (Madhani, 2011). We might think that the newer generation of people might have more edge in this environment when it comes to the current opportunities in technology, but our study shows that this might not be the case. This might be because a lot of other factors comes to play than just an arbitrary factor like an age from which an individual starts to code. Job success will also depend on other essential core skills like communication, leadership, knowledge in other domains, and even the development of other logical skills through other factors (Stack Overflow, 2022). As we also observed the different groups of individuals in the study, it might be better to study each of these groups to get a better understanding of the relationship between the age first started to code and compensation.

Thus, for now, there is an answer for the given research question suggesting that there is no relationship between the age first starting to code and the total compensation of the developer. However, there are still a lot of unanswered questions which was pointed out by this study and could be further explored.

# Future Work

The given data is massive and consists of several clusters of data that should have been isolated and been looked at more in detail to get a more comprehensive story. After this study, it feels like the identified 3 distinct groups of people must be analyzed separately and understand the context-based background they come from.

It seems like a better study will be inspecting the same relation but only taking the data representing the average people who start coding from the age of 5 years to 30 years and ignoring the late learners and early high earners who have skewed the results by a considerable degree.

The other work which can also be done is to investigate these outlier groups from the average developer and understand their context and answer the questions regarding their success from a very young age and regarding the start of coding later in their life without having any problems in their success.

# References

Andrade, C., 2019. The P Value and Statistical Significance: Misunderstandings, Explanations, Challenges, and Alternatives. *Indian J Psychol Med,* 3(41), p. 210–215.

Balkin, D. B. & Gomez-Mejia, . L. R., 1990. Matching Compensation and Organizational Strategies. *Strategic Management Journal,* Volume 11, pp. 153-169.

Cambridge University Press, 1995. *Cambridge international dictionary of English.* Cambridge: Cambridge University Press.

Dalgraad, P., 2008. *Introductory Statistics with R.* 2nd ed. s.l.:Springer.

Herrnstein, R. J. & Murray, C., 1994. *The Bell Curve: Intelligence and Class Structure in American Life.* s.l.:Free Press.

Lund Research Ltd, 2018. *Spearman's Rank-Order Correlation.* [Online]   
Available at: https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php  
[Accessed 7 January 2022].

Madhani, P. M., 2011. Effective Compensation Management for Competitive Advantages: Study of Indian IT Sector. *Journal of Science, Technology and Management,* 3(4), pp. 29-42.

MaRS, 2022. *Employee compensation: Salary, wages, incentives and commissions.* [Online]   
Available at: https://learn.marsdd.com/article/employee-compensation-salary-wages-incentives-and-commissions/

R Core Team, 2021. *R: A Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing.

Spearman, C. E., 1904. *The proof and measurement of association between two things.* Toronto: Robarts - University of Toronto.

Stack Overflow, 2022. *Stack Overflow Annual Developer Survey.* [Online]   
Available at: https://insights.stackoverflow.com/survey/