Benefits of Learning Coding and Machine Learning

Rosy Xu & Jingyu Wu

0 Contribution Statement

Both Rosy Xu and Jingyu Wu wrote R code according to their written parts in this work. Both students contributed equally to the Introduction, Basic Analysis, Advanced Analysis section, and Conclusion. In addition, both students reviewed and added changes to the whole report.

1 Introduction

As we are stepping into the age of artificial intelligence, the job market is having an **increasing demand** for **coding** and **machine learning experiences** among employees. More and more people started to realize this situation and use online sources to learn and practice coding. **Kaggle**, as one of the most widely-used Machine Learning and Data Science communities, conducted an industry-wide survey that presents a comprehensive view of the state of data science and machine learning.

In this project, we utilized the responses of that survey, and generated an analysis of how learning coding and machine learning may be **beneficial** to workers in the aspect of the **level of income and compensations**. We divided responses into groups based on their amount of time spent on coding and machine learning. We used **graphical methods** to compare the distribution of compensations among groups, provided **a point and an interval estimate** of the average amount of compensation for each group, and applied a **regression model** with the time spent on coding and machine learning as explanatory variables and the average compensations as response variable. We were also interested in how **genders correlate** with **compensation** of people in Kaggle. We divided the responses into males and females, and applied a similar **regression model** to each gender to **make comparison** and reveal possible differences. Moreover, we tried to **recommend** several programming **languages** for **new-comers** in data science and machine learning to start with.

2 Section one: will learning coding help with your income?

2.1 Comparison of the distributions of compensation among groups

Method

The columns we picked for this section corresponded to question **6**, **15**, and **24**, which were asking about "time of **writing code**", "time of using **machine learning** methods", and "yearly **compensation**" respectively. Then we did **data cleaning**, and **removed** the responses that were **blank** in question **6** or in question **24**.

After that, we calculated the **estimated time of coding** for each row by **adding** up the response to question **6** and the response to question **15**. For each of the choices in question **6** and question **15**, we used the median of the time strand in the response as our numerical data. For example, if the response is "**3 to 5 years**" for coding time and "**1 to 2 years**" for time of using machine learning methods, then the **estimated time of coding** of that response will be **4 + 1.5 = 5.5** years.

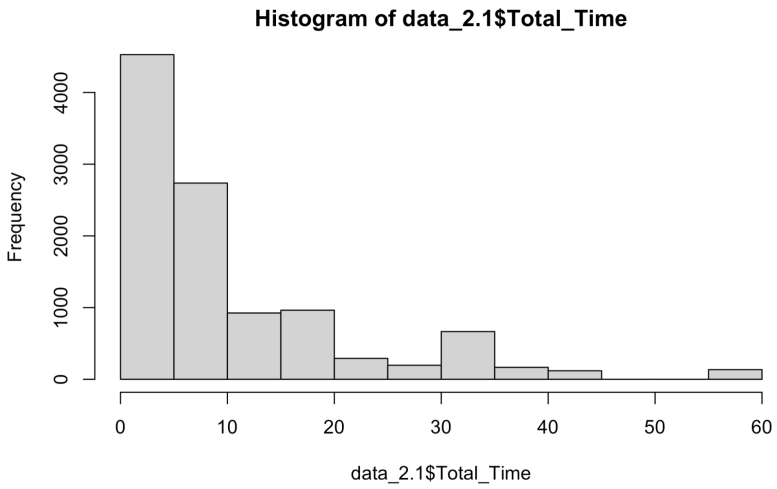
Then we parsed the responses of question **24** and gave each response an **'income level'** by simply following the **grouping** provided by the survey and giving **ascending indices** to them. For example, if one respondent answered "**$0-999**", then it will be grouped to **group 1**, since it is the smallest income group, if one respondent answered "**> $500,000**", then it will be grouped to group **25**, since it is the **25th** smallest group.

Finally, based on the **estimated time of coding**, we divided the responses in groups, and investigated the **distribution** of the **level of income** in each group and the **difference** in **average income** among groups.

Analysis

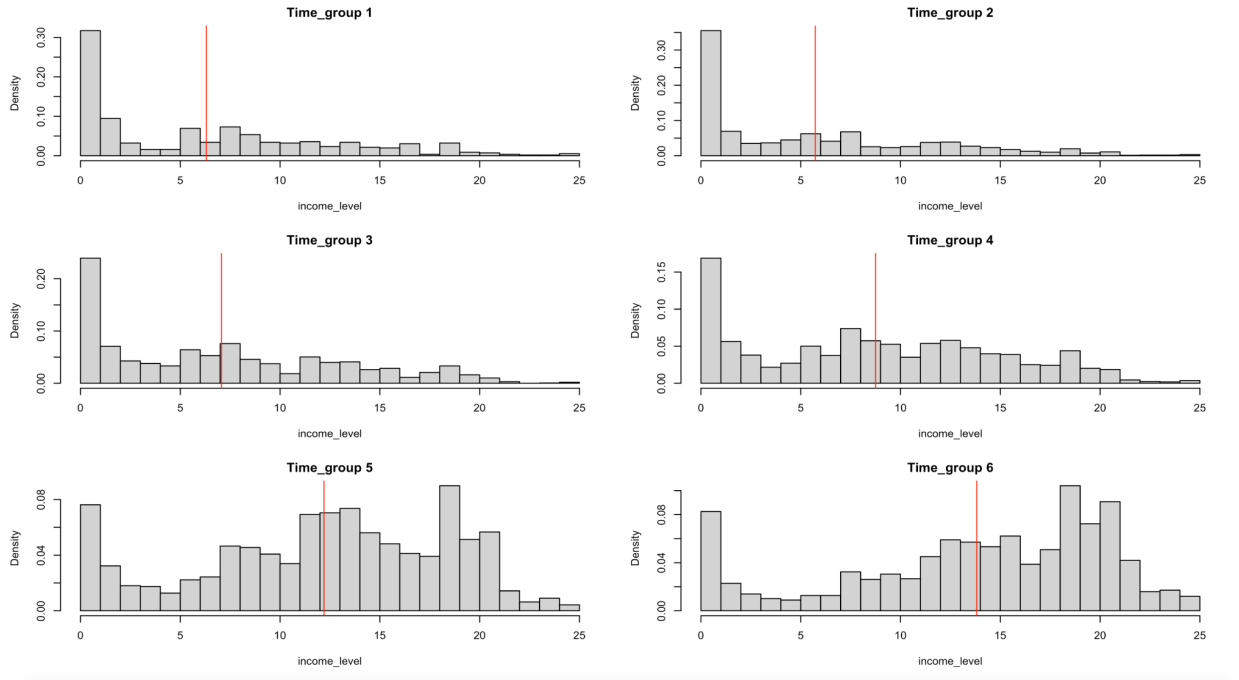
As described above, we calculated the total time of coding of each response, and found the following distribution **(Graph 2.2.1)**. Then we divided it into **6** groups based on the **following standard** (the **distribution of grouping** can be found in **Appendix 1**):

{1: [0,0]; 2: (0,2]; 3: (2,5]; 4:(5,10]; 5: (10,20]; 6: (20,∞)}



Graph 2.1.1 Distribution of Total Time of coding

Then we gave each response an **'income level'** by following the process described above, and we plotted out the **distribution** of income in each group with the **red line** as the **average income level (not the level of average income)**:



Graph 2.1.2 Distribution of income level in each group

We can observe from **Graph 2.1.2** that the distributions in the first three groups were about the same, while there was a clear **increasing trend** in the **occurrence** of **high income levels**, as well as an **increase** in the **average income level**, as the group number increased from **3** to **6.**

Conclusion

We found that the **distributions** of **income** for people learning no or **less than 5 years** of programming was **similar** to each other, which indicates that **little or no experience** in coding and machine learning **may not help much** on your **income**, but as you **learn more** and **deeper** into the topic, there is a **positive correlation** between the **time of coding** and the **level of annual compensation**.

2.2 Point and interval estimate of the average compensation of each group

Method

Similar to the previous part, we still categorized our data into six groups. We calculate the **average** compensation of the group of people who have programming experience for **0 years**, the group of people who have experience for **(0, 2] years**, the group of people who have experience for **(2, 5] years**, the group of people who have experience **(5, 10] years**, the group of people who have experience for **(10,20] years**, and the group of people who have experience for **more than 20 years** respectively to **estimate** the population's average compensation. Then we used the **upper and lower bound** of each compensation level in the choices of question 24 to find the interval estimate.

Analysis

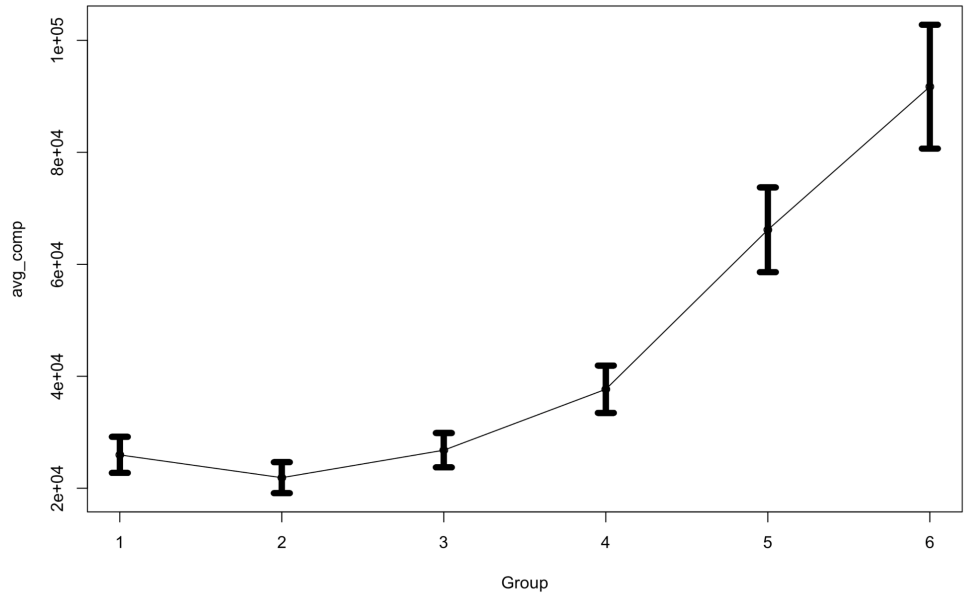
From **Table 2.2.1**, people who have no programming experience have about **25959** dollars **on average** and the **estimated interval** is between **22735** and **29183** dollars. For people in **group 2**, the point estimate of their average compensation decreased to **21883** dollars.While for people in group **3**, the point estimate of average compensation increased back to **26796**. The point estimate of average compensation for **group 4**, **5**, **6** is **37683**, **66166**, and **91728** respectively. The interval

for group **3** is smallest, while the interval for group **6** is largest. This represents that the compensations are relatively **centralized** among group **3** and the compensations **vary a lot** in group **6**.

	Group1	Group2	Group3	Group4	Group5	Group6
point	25959.00	21883.33	26796.24	37683.23	66165.70	91728.89
lower	22735.29	19118.42	23729.40	33457.25	58599.26	80666.03
upper	29182.71	24648.25	29863.07	41909.21	73732.13	102791.75

Table 2.2.1 Point estimate and interval for each group in numeric table

Overall, we could **reconfirm** that compensation stays quite stable when a person starts to learn coding, but the income will increase dramatically when he or she has a longer programming experience from the **graph 2.2.2**.



Graph 2.2.2 Point estimate and interval for each group in line graph

Conclusion

From both the table and the graph, we have seen that even though the estimate of compensation in **group 2 decreases** a bit from **group 1**, there is a **positive relationship overall** between **programming time** and **average compensation**. The result **matches** with the **group estimation** we did in **2.1**, indicating that the **benefits** of **learning coding** will be shown in the long run, and that coding requires a **high level of commitment**, and **quitting early** will only give you **limited return**, which makes your time kind of **wasted**.

2.3 Regression model between compensation and time of learning coding

Method

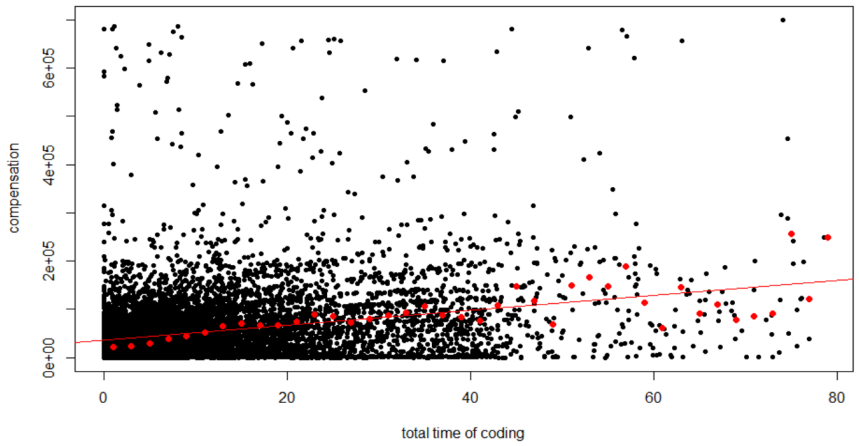
To discover a **numerical correlation** between **annual compensation** and the **time of learning coding**, we made both variables **'continuous'** by assigning them a random value in the group they chose. For example, if the respondent chose "**3 to 5 years,**" we choose a **random real number uniformly** from the range **[3,5]** for it.

After assigning random values to questions **6**, **15**, and **24** to represent their "time of **writing code**", "time of using **machine learning** methods", and "yearly **compensation**" respectively, we **added** the "time of **writing code**" and the "time of using **machine learning** methods" together to represent the **total time of coding**.

Finally, we calculated the **average income** of people learning **[0,2)** years, **[2,4)** years etc., and generated a **linear model** of the averages.

Analysis

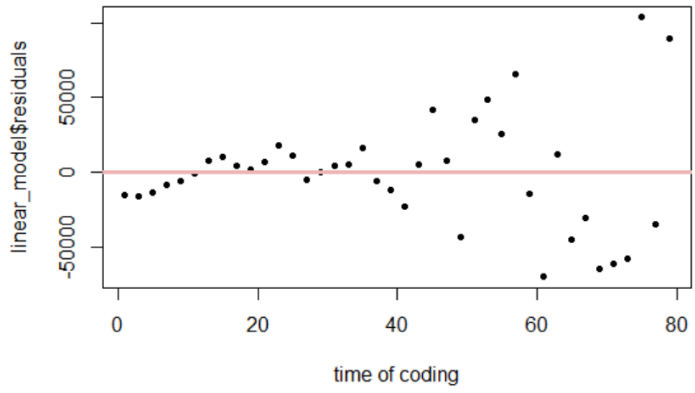
As described above, we used a **random number generator** from a **uniform distribution** to assign values to variables "time of **writing code**", "time of using **machine learning** methods", and "yearly **compensation**" based on the choice of the respondent. Then we **added** the "time of **writing code**" and the "time of using **machine learning** methods" together to represent the **total time of coding**. We plotted out a graph using compensation as **dependent variable** and time of coding as **independent variable**.



Graph 2.3.1 Scatter plot of compensation against time of coding with averages with linear regression line of the averages

After that, we calculated the **average income** of people learning **[0,2)** years, **[2,4)** years etc., and added those points to Graph 2.3.1 in red. We can already observe a slightly increasing trend in average incomes. To verify that, we applied a **linear regression** model to the **averages**, and added the **linear regression line** to the graph. The statistics were "**income($) = 1554.7*(year of learning coding) + 35727.9,**" and **R-squared** equals to **0.487**

Furthermore, we analyzed the **fitness** and **behavior** of our model. From the **residual plot**, we observed that the **residuals** to the **end** of the line are **more variable** than those at the **beginning** of the line, indicating that **homoscedasticity** is **compromised**.



Graph 2.3.2 Residual Plot of the Linear Model

Conclusion

Even though **homoscedasticity** is probably **not achieved** in our data, we still believe, from the scatter plot **(Graph 2.3.1)**, that linear regression line can be **representative** of our data, and is capable of describing the **overall trend**. Based on the linear regression, we conclude that the average income is described by the line "**income($) = 1554.7*(year of learning coding) + 35727.9,**" with **R-squared** equals to **0.487**. Therefore, on average, each year of learning coding

2.4 Is there a gender difference in the annual compensation?

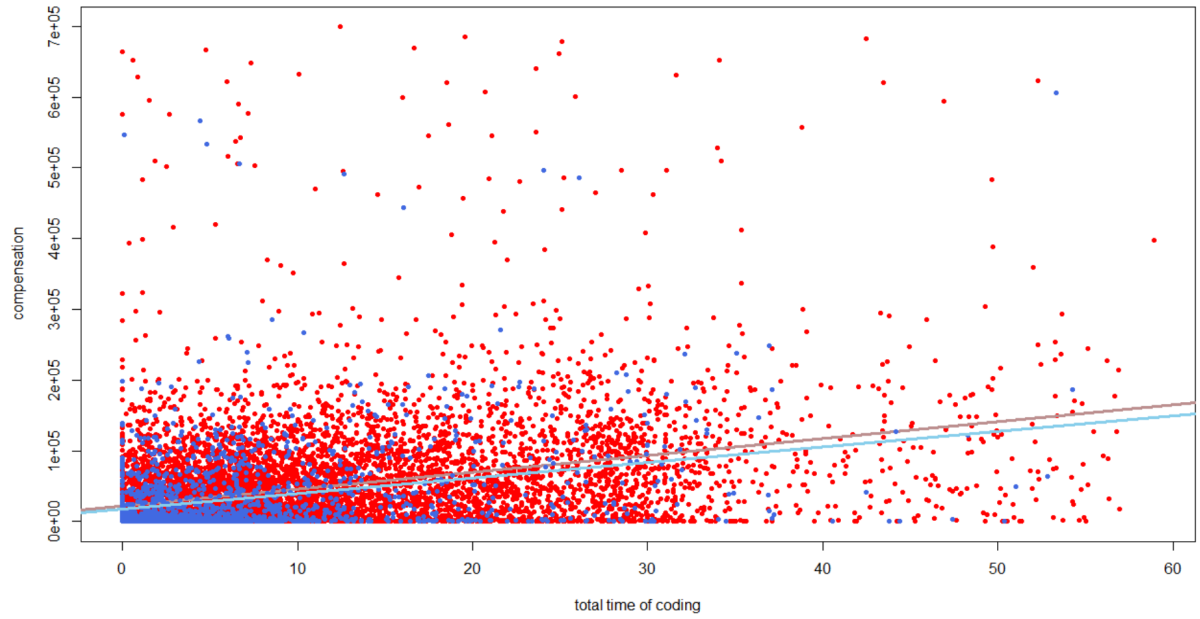
Method

We were also interested in the **possible gender difference** in the career of the field of **coding** and **machine learning**. To investigate that, we started by dividing our data into **groups** of males and females, and applied similar methods from **Section 2.3** by assigning **random values** based on the group each respondent chose, calculating **total coding time**, calculating **average compensation** in each block of **2 years**, and applying **linear regression model** of **compensation against total time of coding** to the data in each gender group.

Analysis

To start with, we divided our data based on the responses to **Question 2** which was asking the gender. We seperated the data into two groups of males and females and **excluded** all other responses. Then we utilized similar data processing techniques as in **Section 2.3**, by assigning **random values** based on the group each respondent chose, calculating **total coding time**, calculating **average compensation** in each block of **2 years**, and applying **linear regression model** of **compensation against total time of coding** to the data in each gender group.

After that, we plotted all data into one graph **(Graph 2.4.1)** to see the difference. The data points in **red** are the data from the **male** group, and the data points in **blue** are the data from the **female** group. The **pink** line represents the **linear regression line** of the averages of **male** group, and the **sky-blue** line represents the linear regression line of the averages of **male** group.



Graph 2.4.1 Scatter plot of data from both males (red) and females (blue) with linear regression lines of averages of both males (pink) and females (sky-blue)

The **male** group has a linear regression line with **y-intercept** of **22282.66**, and a **slope** of **2369.07**, while the **female** group has a linear regression line with **y-intercept** of **17891.6**, and a **slope** of **2192.8**. On average, the female group earns **11%** less than male group.

Conclusion

We conclude, from the graph and the statistics of both regressions, that male workers generally **earn more** than female workers, and will be **earning** even more as time passes by. However, women workers earn about **11%** less than men, which is a closer gap compared to the **general case** that female workers earn **17%** less than male workers (EILEEN PATTEN). It indicates that women are **more competitive** in the area of coding compared with other fields.

3 Advanced Analysis -- programming language suggestions for new learners

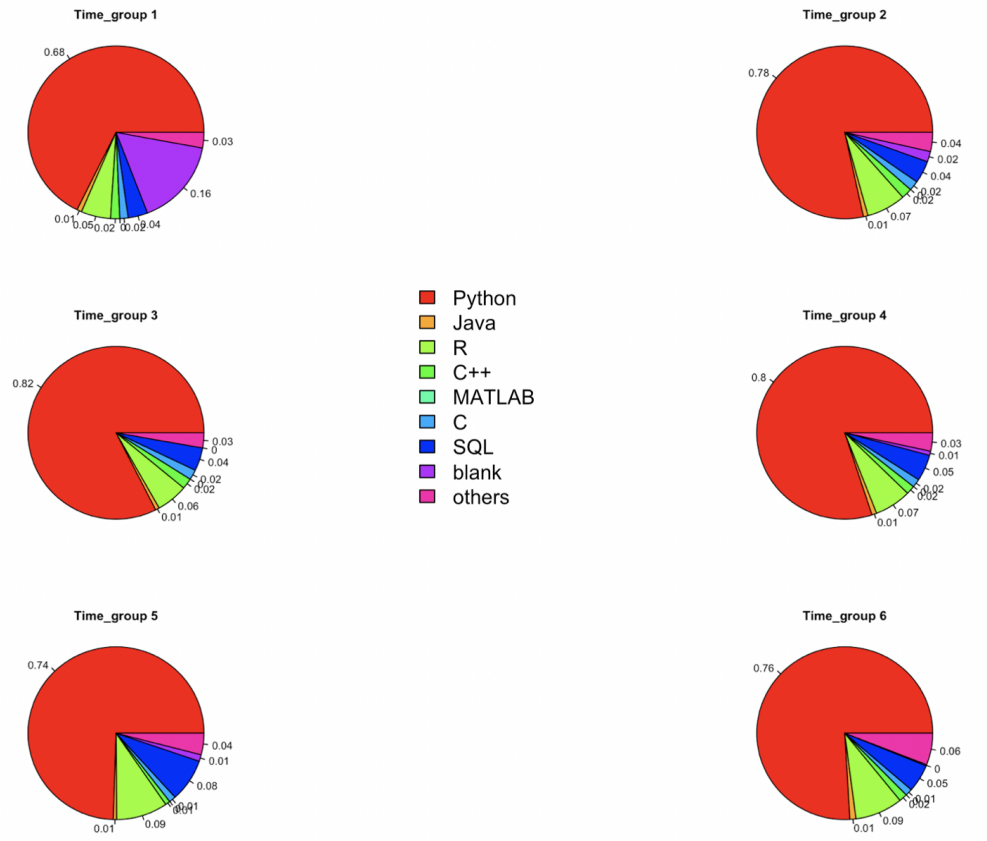
Method

To find the most common programming language suggestions from people who took the survey, we still divide our candidates into 6 different groups. At this time, the 6 groups contain people who have programming experiences for **(0, 2] years**, people who have programming experiences for **(2, 4] years**, people who have programming experiences for **(4, 7] years**, people who have programming experiences for **(7, 10] years**, people who have programming experiences for **(10, 15] years**, and people who have programming experiences for **(15, ∞] years**. We draw **pie graphs** on their answers in question 8, which is "What programming language would you recommend an aspiring data scientist to learn first? After that, we also draw **bar plots** on their answers in question 7, which asks "What programming languages do you use on a regular basis?"

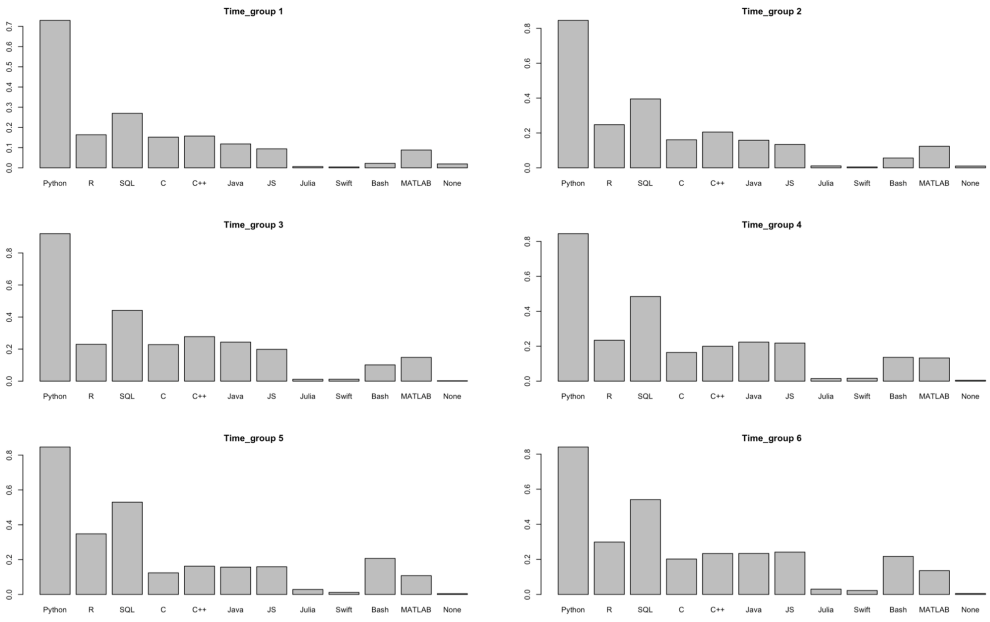
Analysis

From graph 3.1, all six pie graphs indicate that **more than half** of people recommend the incoming data scientists first learn **python**. **R** is the **second highest recommended programming language**. However, the proportion of people who recommend python is almost 14 times of the proportion of people who recommend R, so python is overwhelmingly recommended according to our data. The **third highest recommended programming language** is **SQL**. Even though among the first three groups SQL seems unremarkable, among the last three groups the proportion of people who recommend SQL has approximately the same proportion of those who recommend R. As we learn and really dive into data science, SQL would be more useful than in the beginning.

Similar to the pie graphs, it seems that **most people used R** on a regular basis and the usage of **SQL is more common in the last three groups** in graph 3.2. However, instead of being the third most common programming language, **SQL** is the **second most common programming language** being used in our data. The number of people who use SQL is higher than the number of people who use R. We could infer that R is a rising programming language in recent years, so it has not been used a lot in data science nowadays. But in the future, there will be more and more usages of R and it will replace the current place of SQL in the future. The **third most common programming language** being used is **R.**



Graph 3.1 Pie charts of different groups



Graph 3.2 Bar Plots of different groups

Conclusion

There is no doubt that if people want to dive into data science, the **most recommended programming language** is **python** because it is recommended by more than half of the people in our data. The **second highest recommended programming language** is **R** even though it is not as commonly used as **SQL** on the **current regular basis**. The **third highest recommended programming language** is **SQL**.

5 Conclusion

By dividing our data into **six different groups**, we found that the distributions of compensations of people who have less than 5 years of programming experience were **similar**, while overall, there is a **positive correlation** between people's compensation and the time of programming experiences. We then calculated the **point estimate** and **interval** of the compensations for each group and reconfirmed the positive correlation. We also find the **regression model** of income and time of programming experiences, which is **income($) = 1554.7*(year of learning coding) + 35727.9**. By using the numerical value and graphs, we found that in general, male workers **earn more** than female workers and the gap is **increasing** as the time of programming increases. Furthermore, we used **pie plots** and **bar plots** to give advice on which programming language is most suitable to beginners in data science. The best programming language recommended according to our data is **python.** The second and the third are **R** and **SQL** respectively.

Our research still has several limitations. Even though we have shown that there is a correlation between people's wage and the time of programming experiences, we cannot prove that a longer time of programming causes higher wages by using these data. Also, since our data is from the survey in Kaggle, our sample might be only representative to people who use Kaggle and who want to dive into data science and machine learning.

Works Cited

EILEEN PATTEN. "Racial, gender wage gaps persist in U.S. despite some progress." Pew Research Center JULY 1, 2016

https://www.pewresearch.org/fact-tank/2016/07/01/racial-gender-wage-gaps-persist-in-u-s-despite-some-progress/