**Polish IT Jobs**

Charbel El-Fakhry, ID: 900214262

Asmaa Shaban, ID: 900205076

Joseph Tarbouriech, ID: 900225513

The American University in Cairo

Fall 2022

**Declaration:** We, the three members of the team, worked approximately equally in the project.

1. **Why this data set?**

As all of us in the team are majoring in areas related to the IT world, and out of curiosity, we wanted to work on a dataset that helps us to look over and have an insight of our future; what IT workers use in work? What about the salaries? What are the skills that are demanded? And all other questions that come to our minds as undergraduates. For that, and after a decent research over the datasets, we found many datasets that could help our cause. But, as we are still learning now, we wanted to have recent data that is valuable for our analysis and expected outcome.

Based on what was said, we chose a dataset of Polish IT Jobs in which we saw an opportunity to find valuable findings due to its recent data collection, which was from February 2022 to November 2022, large number of observations ( n = 37787), and variables’ variety which is of number 35.

Through the analysis, we hope to achieve two main targets. First, practice the skills that we have learned in class and apply them in a real world situation. Second, reach a conclusion that helps us in having an idea of the possible future jobs that we might have, and what will help us in getting the best opportunities in the job market.

1. **Questions that can be answered from the data analysis.**

From the data we have, we mainly want to expand the aspects of the jobs and examine the variables that affect the existence of these IT workers in the job market. In addition, we hope to examine the impact of workers’ skills over the salary and employability and skills variables. The questions are around:

1. What is the most common Technological field the employees work in?
2. What is the frequency of employees in the job market according to the experience level?
3. How does permanent employment differ from non-permanent employment?
4. What is the relationship between the skill values and salary in permanent jobs?
5. How does the skill value affect the employability of workers?

1. **Adjusting the data:**

Before analyzing the data, we had to make some adjustments on it. For instance, we replaced the unknown variables in the columns by NA using “Replace” in Excel.

In addition, we had to transform the categorical (nominal) variables into logical ,so we could use them in writing conditional statements in R and in graphs.

There are many variables (35) and we chose to focus on the ones that are more linked to our interest in our future jobs.

Therefore, we chose to analyze in more detail the permanent jobs and their salaries as permanent jobs are the most reliable ones where each one of us seeks to secure one. We decided to ignore the other types of jobs because they aren’t strongly related to our targets and we would like to focus on one area and analyze it thoroughly. By that, we are not only embracing the value of the dataset, but we are also showing our strong planning over the steps in the project.

1. **Description of the data:**

The source that we have used to find the dataset is Kaggle.

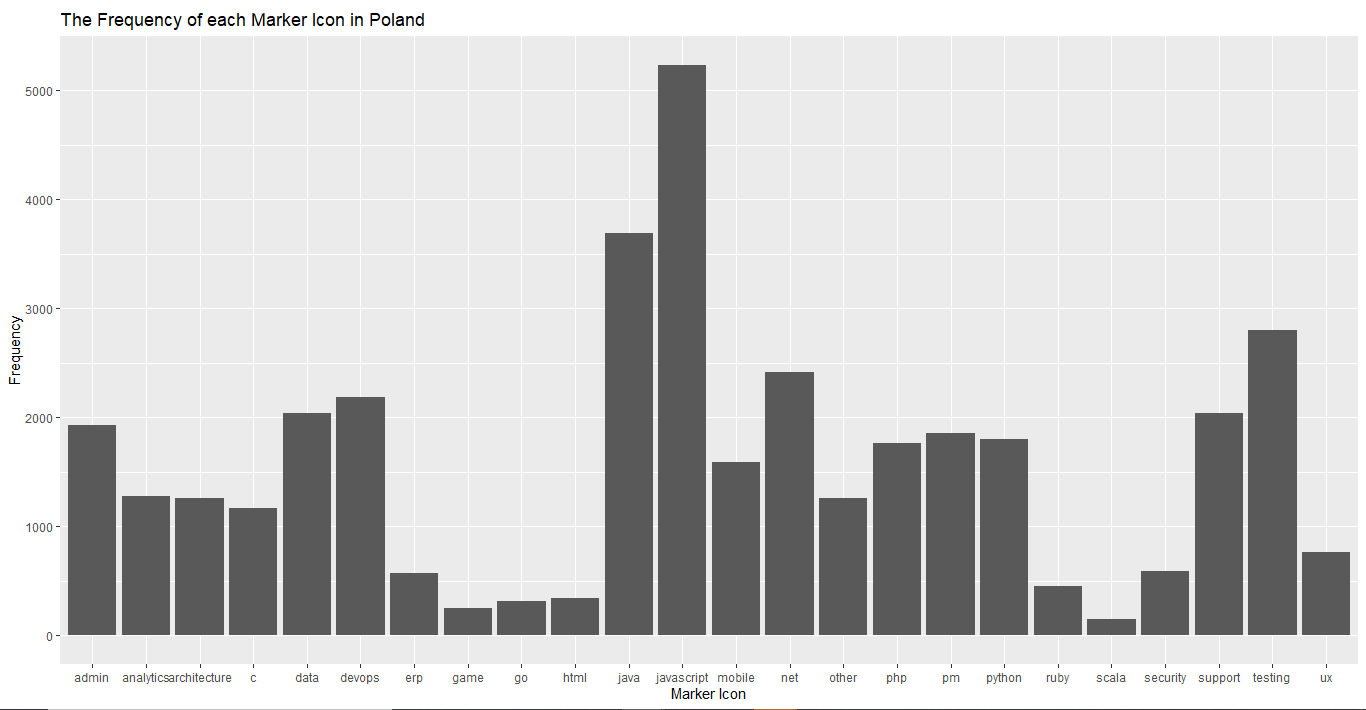
The link for the data set is:

[**https://www.kaggle.com/datasets/kriegsmaschine/polish-it-job-board-data-from-2022**](https://www.kaggle.com/datasets/kriegsmaschine/polish-it-job-board-data-from-2022)

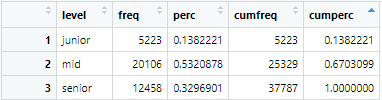
1. **Table defining each of the 35 variables of our dataset:**

| No. | Variable Name | Type | Unit of measurement | Description |
| --- | --- | --- | --- | --- |
| 1 | Title | nominal | NA | Name of the position each employee holds at work. |
| 2 | City | nominal | NA | The city that the employee works from or in. |
| 3 | Country\_code | nominal | NA | The abbreviation of the country from which the employee works in. |
| 4 | Marker\_icon | nominal | NA | The coding language that the employees use in work. |
| 5 | Workplace\_type | categorical | Work type | What is the type that workers are working in? Remotely, partially remotely, or in office? |
| 6 | Experience\_level | categorical | experience | The experience level of the employee which is based on the level of skills, concepts understanding and work time. |
| 7 | Published\_at | ordinal | date & time | The date of when the data was collected following the ISO-8601 date representation. |
| 8 | Remote\_interview | logical | NA | Was the interview made remotely or not? |
| 9 | Remote | logical | NA | This logical variable matches the workplace where if it is remotely, it gives true and if it is partially remotely or in office it gives false. |
| 10 | Open\_to\_hire\_Ukrainians | logical | NA | Is the company open to hiring Ukrainians or not ? |
| 11 | Company\_size\_from | numerical | Number of employees | The size of the company by the number of employees. |
| 12 | Company\_size\_to | numerical | Number of employees | If the size of the company is defined, the value in this variable is 0. But, in the case of not knowing the exact number of employees, there is a range from the previous variable to this variable. |
| 13 | if\_permanent | logical | NA | If the employee is working in a permanent job or not. |
| 14 | salary\_from\_permanent | numerical | Currency | The lower limit of salary range that the employee could take in the case of working in a permanent job. |
| 15 | salary\_to\_permanent | numerical | Currency | The higher limit of salary range that the employee could take in the case of working in a permanent job. |
| 16 | salary\_currency\_permanent | nominal | NA | The currency of the salaries that employees in permanent jobs earn. |
| 17 | if\_b2b | logical | NA | Is the employee working in a company following a business to business strategy ? |
| 18 | salary\_from\_b2b | numerical | Currency | The lower limit of salary range that the employee could take in the case of working in a job for a company following a b2b strategy. |
| 19 | salary\_to\_b2b | numerical | Currency | The higher limit of salary range that the employee could take in the case of working in a job for a company following a b2b strategy. |
| 20 | salary\_currency\_b2b | nominal | NA | The currency of the salaries that employees working in a job for a company following a b2b strategy earn. |
| 21 | if\_mandate | logical | NA | Is the employee working in a company that is mandated or not, so is the allocation of funds to an investment manager managed for a specific purpose or style or not ? |
| 22 | salary\_from\_mandate | numerical | Currency | The lower limit of salary range that the employee could take in the case of working in a mandated job. |
| 23 | salary\_to\_mandate | numerical | Currency | The higher limit of salary range that the employee could take in the case of working in a mandated job. |
| 24 | salary\_currency\_mandate | nominal | NA | The currency of the salaries that employees in mandated jobs earn. |
| 25 | if\_other | logical | NA | Is the employee working in another company, or does he/she have a different job? |
| 26 | salary\_from\_other | numerical | currency | The lower limit of salary range that the employee could take in the case of another job. |
| 27 | salary\_to\_other | numerical | currency | The higher limit of salary range that the employee could take in the case of another job. |
| 28 | salary\_currency\_other | nominal | NA | The currency of the salaries that employees in other jobs earn. |
| 29 | currency\_exchange\_rate | numerical | Original unit to pln | The exchange rate from the Polish Zloty(pln) to the currency that the employee gets his/her salary with. |
| 30 | skills\_name\_0 | nominal | NA | The name of the first skill of an employee |
| 31 | skills\_value\_0 | ordinal | Score out of 5 | The score out of 5 of the first skill of an employee (1:beginner → 5:advanced) |
| 32 | skills\_name\_1 | nominal | NA | The name of the second skill of an employee |
| 33 | skills\_value\_1 | ordinal | Score out of 5 | The score out of 5 of the second skill of an employee (1:beginner → 5:advanced) |
| 34 | skills\_name\_2 | nominal | NA | The name of the third skill of an employee |
| 35 | skills\_value\_2 | ordinal | Score out of 5 | The score out of 5 of the third skill of an employee (1:beginner → 5:advanced) |

1. **Bar chart on the Frequency of employees over the maker Icon (Job work)**

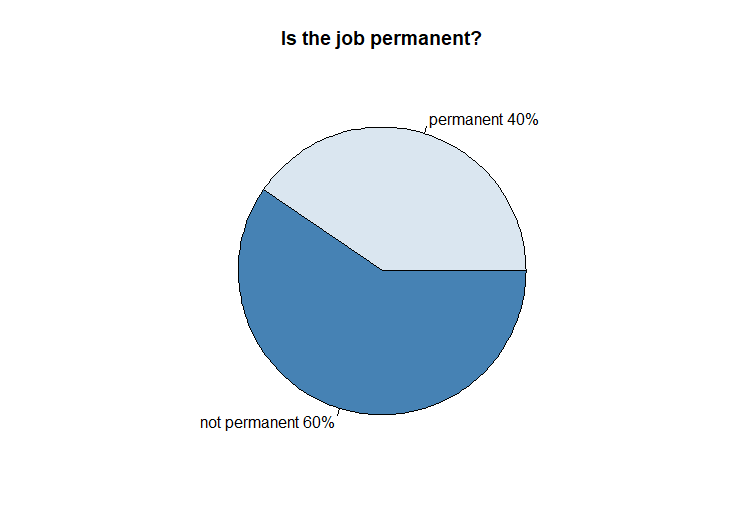
This bar chart shows the frequency of employees over each marker icon which is the type of work or the programming language that Polish IT employees use. As it is seen, the highest frequency of workers were working with **JavaScript.** The least frequency of workers were working with **scala.** The other jobs with relatively high workers’ frequency were working with **java,**  in software **testing, net, devops, support, and data.** The value of that graph is that it reflects on the job market’s demand for IT jobs.

1. **Frequency table of the experience level of workers over their frequency.**



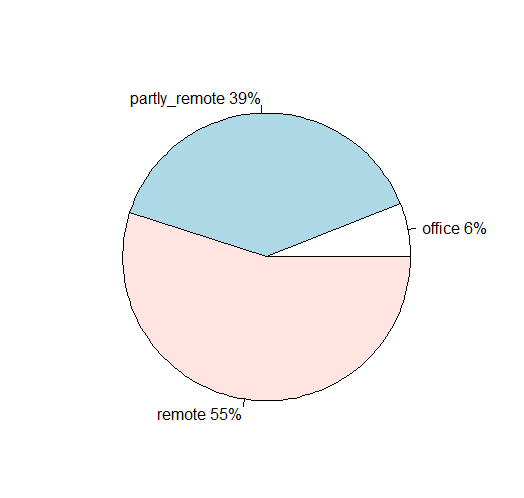
This frequency table distribution shows the frequency of Polish IT workers over their experience level. This can give an insight on the experience level that the job market is mostly looking for. More than 0.5 of the employees are mid and it is the highest percentage of all employees. The least number is the number of juniors.

1. **Pie chart of whether a job is permanent or not**



This pie chart shows the type of jobs that IT Polish employees are in. Specifically, whether the Jobs are permanent or not. It is seen that **40%** of them are permanent employees and **60%** are not permanent employees.

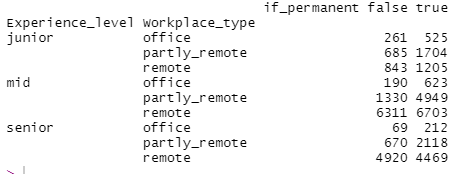
1. **Pie chart of the workplace of employees who work in permanent jobs.**



This pie chart shows the workplace that Polish IT permanent workers work from. It is seen from the graph that **55%** of them work remotely, **39%** partly remotely, and **6%** in the office. From that we can interpret the common workplace of the Polish IT workers , who work in permanent jobs, which is remotely.

**Note:** The data is recent and COVID-19 could explain the reason for the high percentages of workers who work remotely.

1. **Contingency table of how experience level and workplace type of employees impacts if a job is permanent (true) or non-permanent (false)**



This contingency table shows the experience level and workplace\_type of employees in permanent and non-permanent jobs.

This table is not a summary. It shows the data is an organized and more concise way. For the permanent employees, most of them who work in office are mid and junior experience level employees. Most of them who work with the hybrid system(partly\_remote)are mid level employees. And, most of them who work completely remotely are mid level workers.

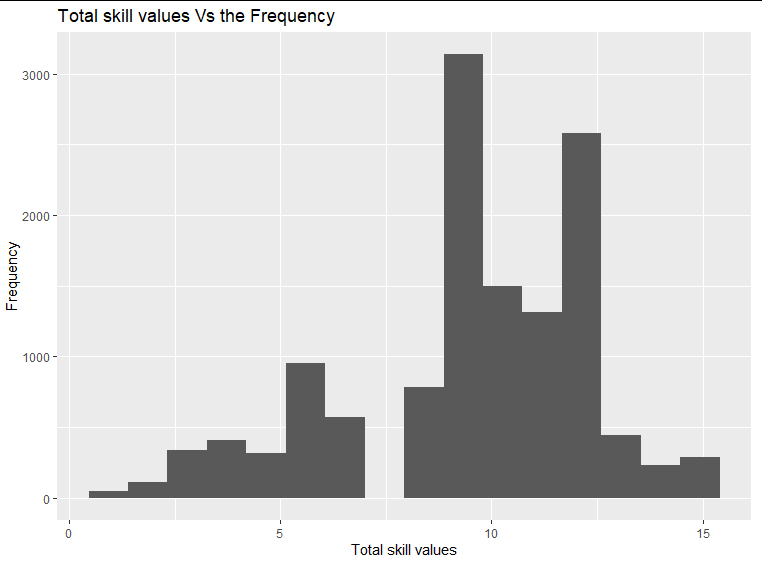
1. **Table showing the numerical analysis of seniors, mids and juniors who have a permanent job.**

|  | **Seniors** | **Mids** | **Juniors** |
| --- | --- | --- | --- |
| **Range of salary from permanent** | 409862 | 372240 | 41818.6 |
| **Range of salary to permanent** | 489697 | 419481.8 | 84617 |
| **Average of salary from permanent** | 17510.97 | 11579.39 | 5815.768 |
| **Average of salary to permanent** | 24314.29 | 17074.18 | 8606.684 |
| **Variance in salary from permanent** | 114815846 | 72155330 | 6214889 |
| **Variance in salary to permanent** | 173913602 | 116079106 | 13575506 |
| **Standard Deviation in salary from permanent** | 10715.22 | 8494.429 | 2492.968 |
| **Standard Deviation in salary to permanent** | 13187.63 | 10774 | 3684.495 |

* From the average salary from/to of permanent workers, we can notice that the highest salary average is for seniors. On the other hand, the lowest salary average is for juniors. From that, we can interpret that the salary increases with work experience level.
* The interpretations made from the average, are similar also with the variance which tells the degree of spread in the data set.
* From the standard deviation, it is noticed that the standard deviation for seniors is larger than the standard deviation for mids, which by its turn is larger than the standard deviation of the salary for juniors’ workers.

This shows that the salaries for permanent seniors are more spread out than the salaries of mids and juniors. By that, juniors’ salaries are more clustered around the mean.

1. **Histogram of the total skill value over the frequency.**



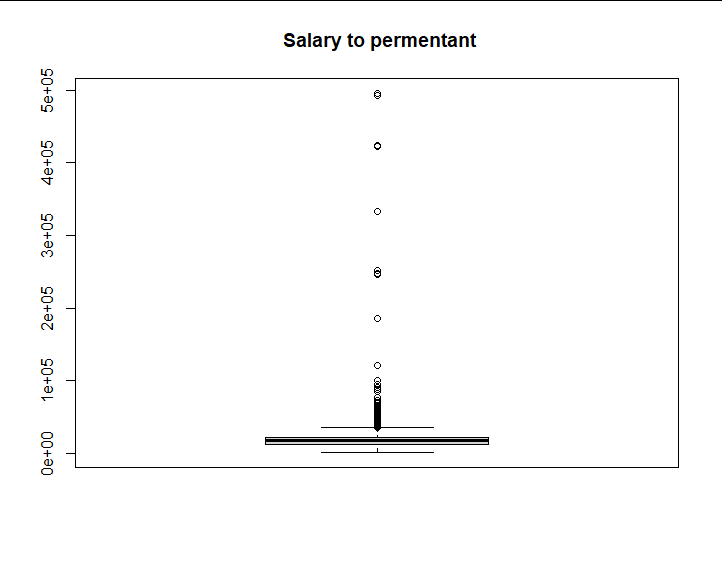
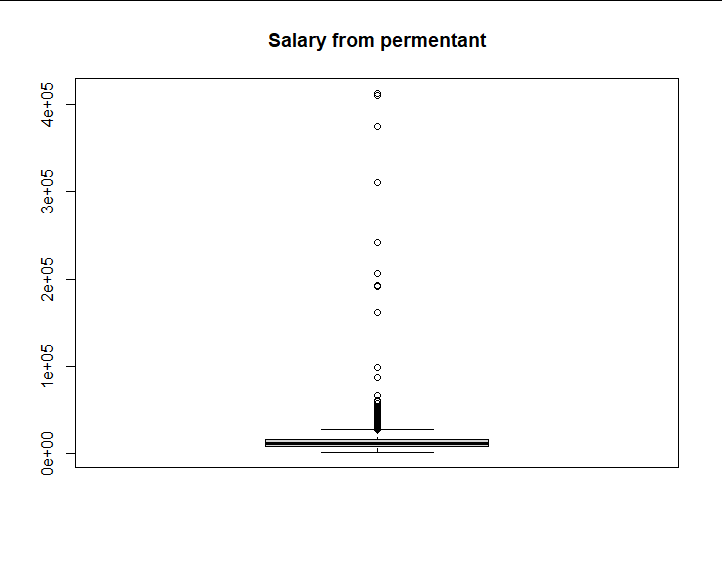
This histogram shows the frequency of employees over the total skill level. The highest frequency of employees is with the ones who have 9 as the total skill value. The least frequency is with employees who have a total skill value 1. From this we can conclude that most employable IT workers are the ones who have at least two high value skills. People with very low or very high skills’ values were not frequent as much in the job market.

1. **Table on the average of the total value of the three skills based on experience level**

| **Experience Level** | **Senior** | **Mid** | **Junior** |
| --- | --- | --- | --- |
| **Average of total value of the three skills** | 10.95479 | 9.219964 | 6.113917 |

From this table, it can be interpreted that seniors are the most skilled between the workers, then mids, then juniors. These numbers are logical where skills are acquired and enhanced with work experience.

1. **Box plots of the five number summary of the salary from/to permanent**



They show that the first, second and the third quartile are closed. Also , the two plots are skewed to the right. The maximum value in “Salary From Permanent” is 413115 and minimum value is 1445.

In the second plot , which represents the “Salary To Permanent”, the maximum salary is 495738 and the minimum salary is 1445. However, the maximum salary in the two plots are outliers.

And there are the six number summary of the two variables:

**Six Number summary:**

1. **Salary from permanent**

Min. 1st Qu. Median Mean 3rd Qu Max

1445 8000 12000 12866 16000 413115

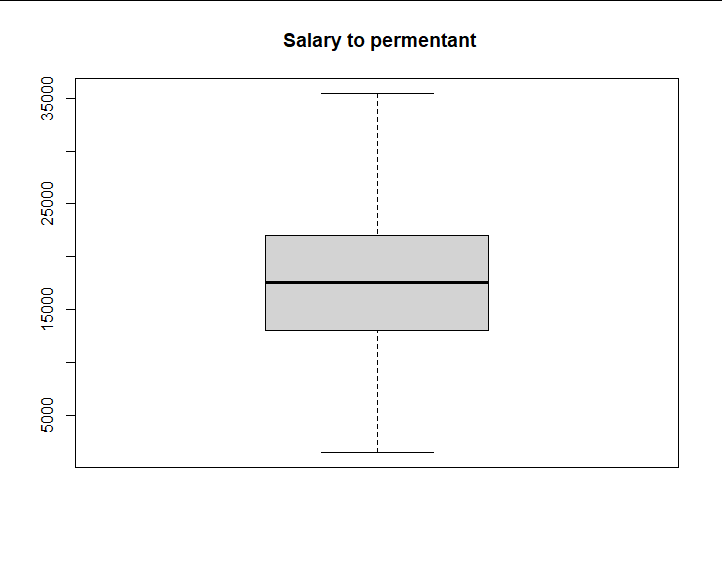
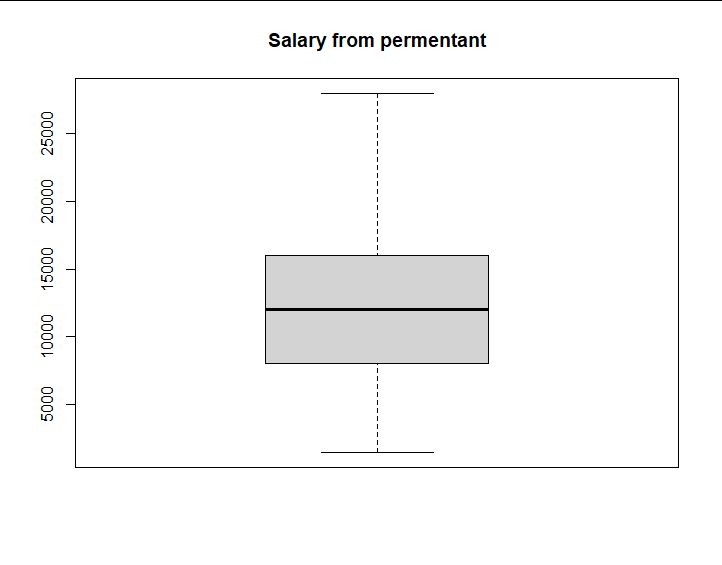
1. **Salary to permanent**

Min 1st Qu. Median Mean 3rd Qu Max.

1445 13000 17600 18448 22000 495738

The previous two numerical summaries, shows us the six number summaries of the salary range(from/to) of the permanent jobs. We can interpret that the data is skewed since there is a difference between the median and mean. In addition, the range from/to permanent salaries is very wide which shows us the variety of observations that we have in the dataset.

1. **Box plots of the salaries after trimming the outliers**

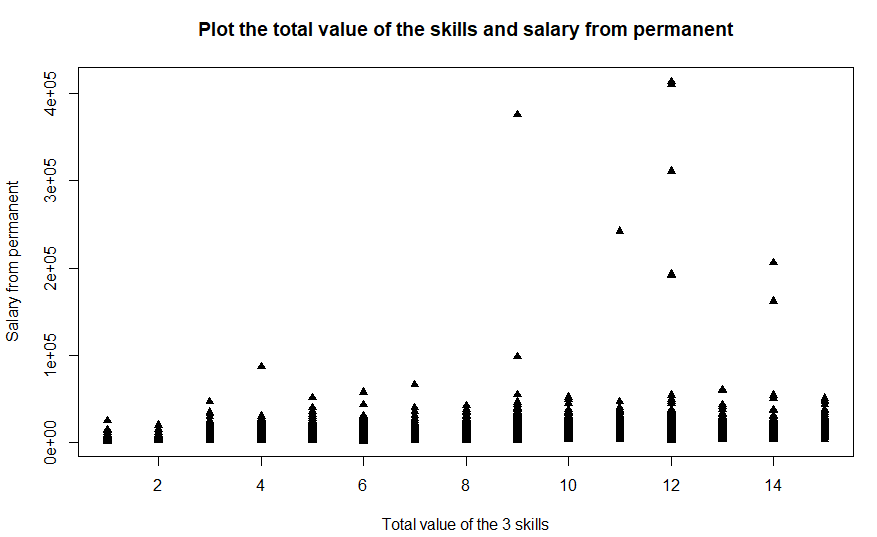
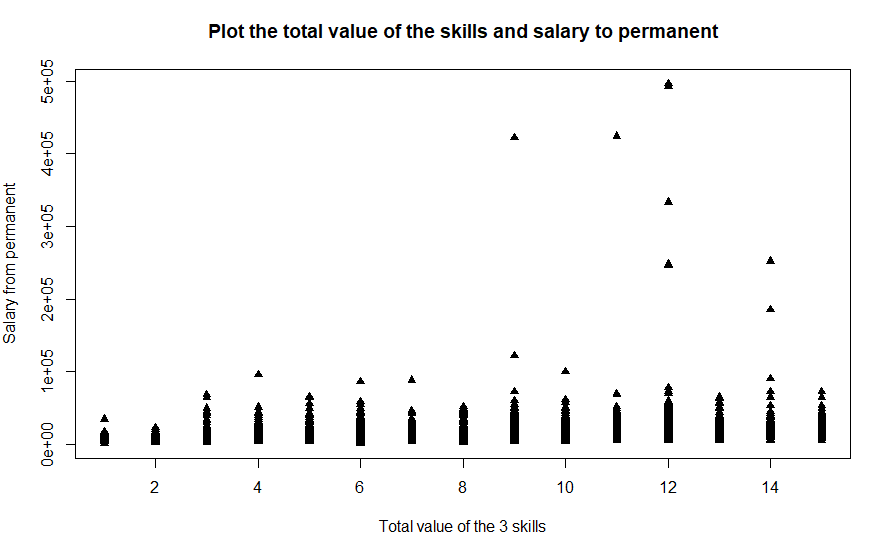
The two box plots represent the “Salary From Permanent” and “ Salary To Permanent”where the outliers are trimmed.

**The first box plot** , “Salary From Permanent” is skewed to the right “positive”.

The minimum value does not change, but the maximum value is between 25000 and 30000.

**In the second plot**, “Salary To Permanent”, the is in the middle, so it is not skewed. The minimum value does not change and the maximum value is about 35000.

1. **Scattered plots of the relationship between the total value of the three skills and salaries.**

****

From the two previous dot plots, which shows the relation between the total value of the skills and salary from/to permanent, it can be interpreted that people with low total value of skills are getting paid less than those who have high skills level. However, the difference in salaries between workers who have a considerable total value skills( 3 or more) and workers with high skills level is minimal and the relationship is not linear after value of 7.

In the graphs, there are some outliers that show some very high salaries in comparison with the group itself. These outliers aren’t understandable inside the frame of the graphs. However, other variables could have an impact on their existence as the experience level.

1. **Conclusion**

In conclusion, and after analyzing and graphing our dataset, we have found that the most common technological field that the employees in the IT field work in is Javascript and Java Coding. Moreover, the most employees in the labor market are mid and seniors which shows us the positive correlation between experience level and employability. Furthermore, based on the contingency table, we have found out that most jobs rely on permanent workers due to their larger frequency in comparison to non-permanent workers. Then, from the scattered plots, we observed that the more skillful the employees are, the larger their salaries are. However, the differences are not major. Moreover, based on the histogram, we found out that skilful employees (with at least 2 job-related skills) have a better chance of getting employed than less-skilled employees. Also, we have found that employees with a high value of skills have an edge to earn more than those with less value. Finally, this project is not just an application of our knowledge, but a lesson that introduced us to a real life environment, taught us how to work in a team, and made us more excited to discover the world of DATA!