Looking into each and every variable individually for a detailed analysis. Since this data is divided into categories, we can do the Descriptive Statistics and analyze the data from the Data Profiling report.

Plotting the histograms of each variable along with the qq plots, we can check the frequency distribution of each variable.

We need not check the candidate ID variable

The **Basic Statistics** shows us the rows and columns which are continuous and discrete. If there are any missing columns and observations in the entire data. The total number of records and observations along with the memory allocated to the entire dataset.

Percentages shows us the graphical representation of the above basic statistics data.

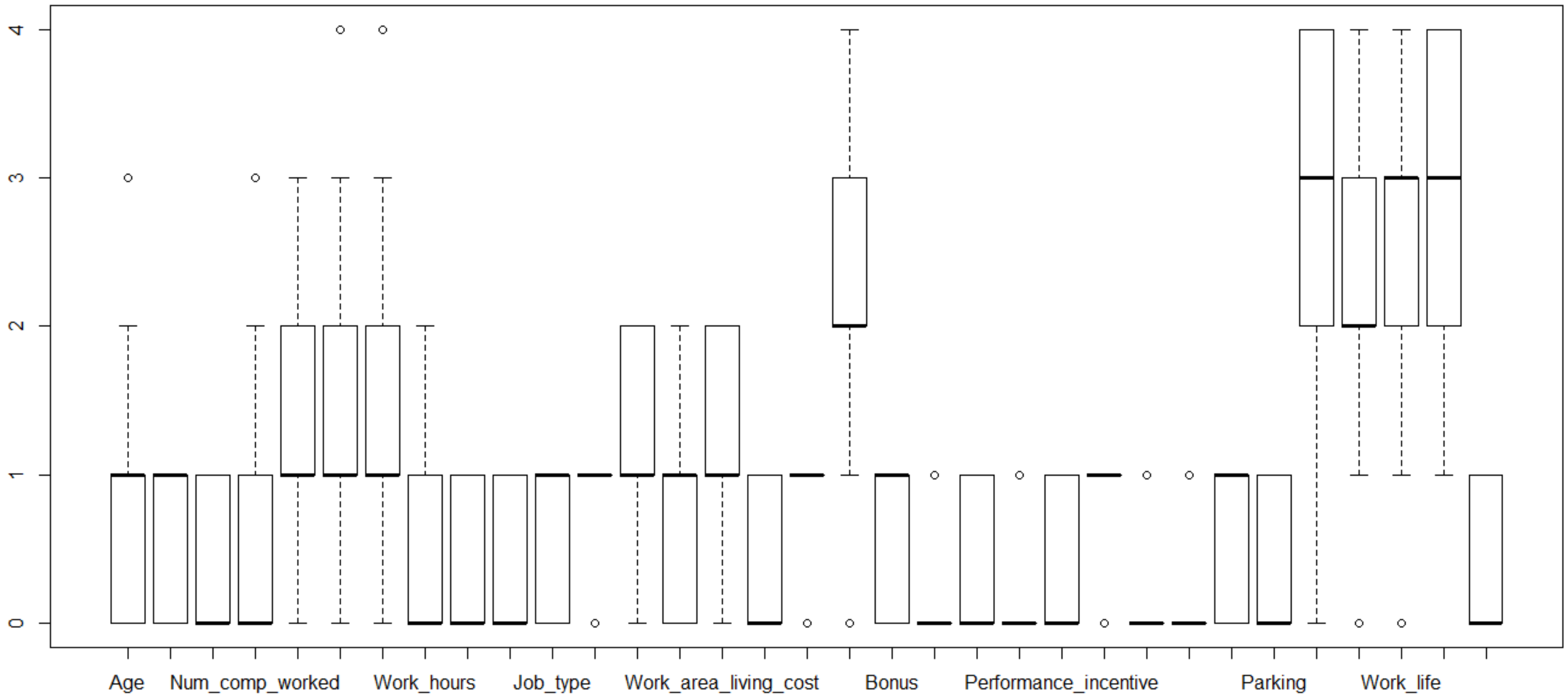
Data structure shows a mind map representation of the data where which variables are there in the total dataset.

Missing Data shows us what percentage f data is missing in each variable. Since it has n missing data it is showing 0% for every variable.

Univariate Distribution shows us the Histograms for each variable.

QQ Plot Shows us the variables distribution and linearity. All the 33 variables are plotted by the qq line and plot where we can check the linearity of the variables. Since it is a categorical and continuous data, we can clearly see that the data variables are in the visualized as step from r in a binary format. Normal distribution cannot be seen in this variable.

Correlation Analysis is the major part where we can find the exact relation between each variable with the other in the data. We can clearly see that there are certain variables which are highly dependent on each other and showing more than 75% correlation.



Above is the box plot for all the variables. We can see the outliers from the following variables : Age, Num companies worked, last salary, salary hike, company possess vehicle, current employment, offer letter process time, allowance, employee share scheme, annual leave, medical family, child day care, comp rating and work life.

Using the data explorer package, we can have the visualization of each variable individually how it is distributed and also how are the qq plots are placed.

Below we can have the distribution plots along with the qq plots for their linearity check.

Basic Statistics Raw Counts **Name**	**Value**
Rows	3,527
Columns	33
Discrete columns	0
Continuous columns	33
All missing columns	0
Missing observations	0
Complete Rows	3,527
Total observations	116,391
Memory allocation	462.4 Kb

The ID Column was omitted since we have no use of that into our analysis. As per our previous analysis, there are 1712 missing values in the Nursing room column and they were replaced using the imputation method. We used the Mice and VIM packages for the imputation and re written the excel with the variable.

Business Problem:

If the candidate offered will accept it or not. In the company interviews, all the candidates who got offers and their acceptance or rejected offer data is been given to us and that is including all the variables which are mentioned as above.

Out of these, there are certain variables which can be directly correlated with the dependent variable. On that case, first we need to find the correlation coefficients or the strength of correlation between each variable with the dependent variable as well as with one another to check the heteroscedasticity.

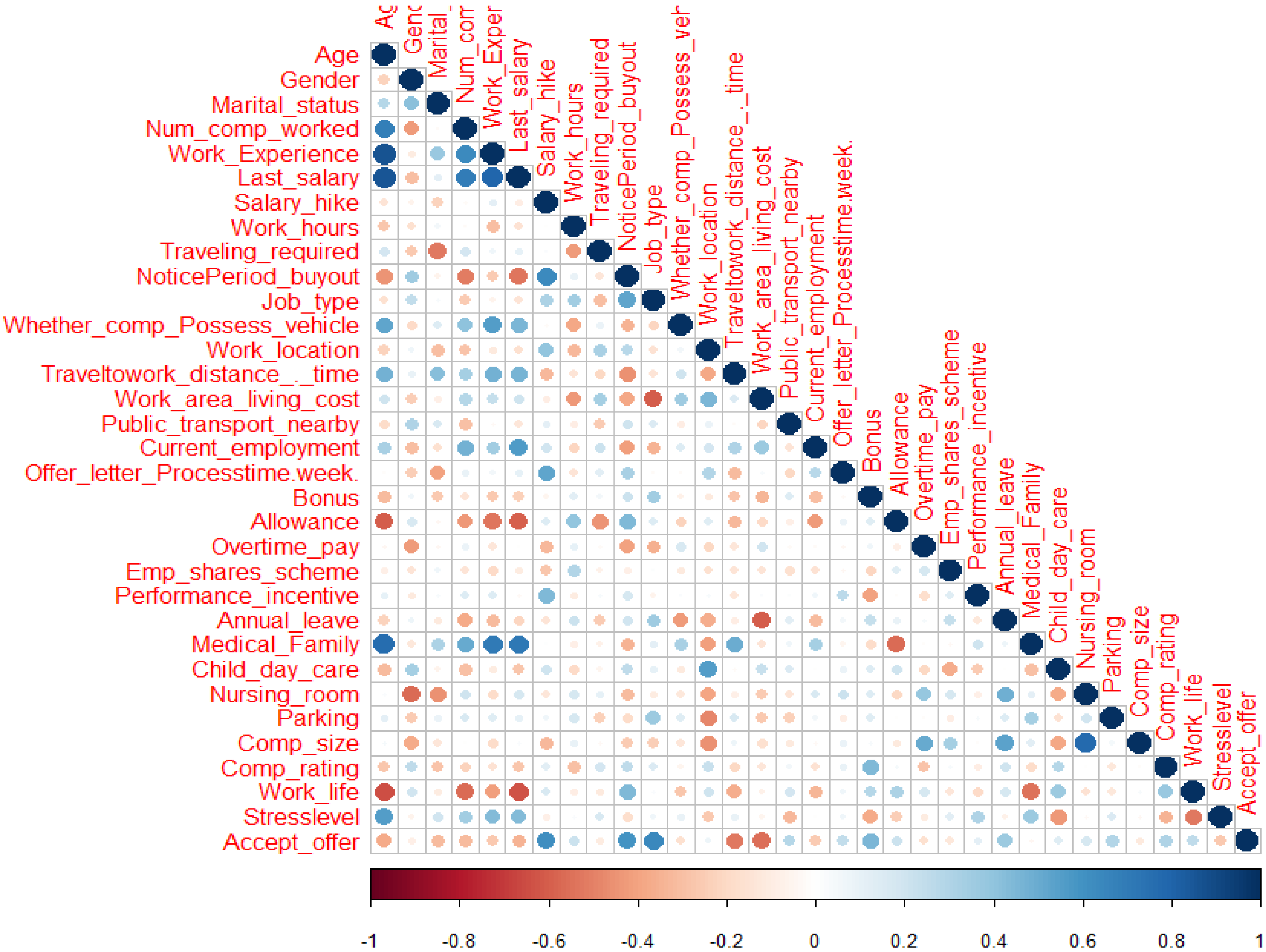
Looking into the summary of the variables in a detailed way below.

```
n obs: 3527
n variables: 33

-- Variable type:integer -----------------------------------------------------------
                         variable missing complete    n  mean    sd p0 p25 p50 p75 p100     hist
                     Accept_offer       0     3527 3527 0.48   0.5   0   0   0   1    1 ▇▁▁▁▁▁▁▇
                              Age       0     3527 3527 0.8    0.77  0   0   1   1    3 ▇▁▇▁▁▃▁▁
                        Allowance       0     3527 3527 0.2    0.4   0   0   0   0    1 ▇▁▁▁▁▁▁▂
                     Annual_leave       0     3527 3527 0.86   0.35  0   1   1   1    1 ▁▁▁▁▁▁▁▇
                            Bonus       0     3527 3527 0.52   0.5   0   0   1   1    1 ▇▁▁▁▁▁▁▇
                   Child_day_care       0     3527 3527 0.094  0.29  0   0   0   0    1 ▇▁▁▁▁▁▁▁
                      Comp_rating       0     3527 3527 2.38   1.02  0   2   2   3    4 ▁▂▁▇▁▅▁▂
                        Comp_size       0     3527 3527 2.76   1.24  0   2   3   4    4 ▂▁▁▂▁▇▁▇
               Current_employment       0     3527 3527 0.83   0.37  0   1   1   1    1 ▂▁▁▁▁▁▁▇
                Emp_shares_scheme       0     3527 3527 0.12   0.33  0   0   0   0    1 ▇▁▁▁▁▁▁▁
                           Gender       0     3527 3527 0.54   0.5   0   0   1   1    1 ▇▁▁▁▁▁▁▇
                         Job_type       0     3527 3527 0.62   0.48  0   0   1   1    1 ▅▁▁▁▁▁▁▇
                      Last_salary       0     3527 3527 1.38   0.98  0   1   1   2    4 ▃▇▁▃▁▁▁▁
                   Marital_status       0     3527 3527 0.44   0.5   0   0   0   1    1 ▇▁▁▁▁▁▁▆
                   Medical_Family       0     3527 3527 0.18   0.38  0   0   0   0    1 ▇▁▁▁▁▁▁▂
               NoticePeriod_buyout       0     3527 3527 0.36   0.48  0   0   0   1    1 ▇▁▁▁▁▁▁▅
                  Num_comp_worked       0     3527 3527 0.56   0.87  0   0   0   1    3 ▇▁▃▁▁▁▁▁
                     Nursing_room       0     3527 3527 0.57   0.5   0   0   1   1    1 ▆▁▁▁▁▁▁▇
 Offer_letter_Processtime.week.       0     3527 3527 2.19   0.9   0   2   2   3    4 ▁▂▁▇▁▂▁▂
                     Overtime_pay       0     3527 3527 0.26   0.44  0   0   0   1    1 ▇▁▁▁▁▁▁▃
                          Parking       0     3527 3527 0.48   0.5   0   0   0   1    1 ▇▁▁▁▁▁▁▇
             Performance_incentive       0     3527 3527 0.28   0.45  0   0   0   1    1 ▇▁▁▁▁▁▁▃
           Public_transport_nearby       0     3527 3527 0.47   0.5   0   0   0   1    1 ▇▁▁▁▁▁▁▇
                      Salary_hike       0     3527 3527 1.25   0.95  0   1   1   2    4 ▃▇▁▃▁▁▁▁
                      Stresslevel       0     3527 3527 2.85   0.99  1   2   3   4    4 ▃▁▆▁▁▇▁▇
                Traveling_required       0     3527 3527 0.49   0.5   0   0   0   1    1 ▇▁▁▁▁▁▁▇
       Traveltowork_distance_._time       0     3527 3527 0.78   0.73  0   0   1   1    2 ▇▁▁▇▁▁▁▃
       Whether_comp_Possess_vehicle       0     3527 3527 0.85   0.36  0   1   1   1    1 ▂▁▁▁▁▁▁▇
            Work_area_living_cost       0     3527 3527 1.2    0.72  0   1   1   2    2 ▃▁▁▇▁▁▁▇
                  Work_Experience       0     3527 3527 1.3    0.89  0   1   1   2    3 ▃▁▇▁▁▆▁▂
                       Work_hours       0     3527 3527 0.54   0.71  0   0   0   1    2 ▇▁▁▃▁▁▁▂
                        Work_life       0     3527 3527 2.5    1.16  0   2   3   3    4 ▂▁▁▃▁▇▁▂
                    Work_location       0     3527 3527 1.05   0.68  0   1   1   2    2 ▂▁▁▇▁▁▁▂
```

The above is a tabular representation of the box plots with the values given and the histograms plotted beside.

Since there are 33 variables and we need a 33*33 table for the correlation matrix, we can check it pictorially as follows.



From the above correlation visualization, we can easily point out which variables are having more correlation and which are not.

As per the above diagram, the dependent variable not highly correlated either positively or negatively to any of the variable. The maximum value is less than 50% correlation and it may not be taken into consideration.

Variable 1	Variable 2	Correlation
Age	work experience	0.62
Age	last salary	0.56
Age	medical family	0.51

Only these three variables exhibit more than 50% correlation with the Age variable. Since the values are less we can consider those variables and state that there is no heteroscedasticity.

Doing a regression analysis on the data to check how the variables react to the dependent variable.

```
Call:
glm(formula = Accept_offer ~ ., family = "binomial", data = train)

Deviance Residuals:
Min       1Q   Median       3Q      Max
-2.3172  -0.0666  -0.0001   0.0744   4.5102

Coefficients:
Estimate Std. Error z value             Pr(>|z|)
(Intercept)                    -17.18239    1.75218  -9.806 < 0.0000000000000002 ***
Age                              4.32057    0.41191  10.489 < 0.0000000000000002 ***
Gender                          -4.54828    0.50010  -9.095 < 0.0000000000000002 ***
Marital_status                   0.93657    0.44312   2.114              0.03455 *
Num_comp_worked                 -3.10353    0.36094  -8.598 < 0.0000000000000002 ***
Work_Experience                 -2.67265    0.34568  -7.732  0.00000000000001062 ***
Last_salary                     -3.53966    0.30074 -11.770 < 0.0000000000000002 ***
Salary_hike                      3.38436    0.34492   9.812 < 0.0000000000000002 ***
Work_hours                       1.17299    0.23504   4.991  0.00000060161359316 ***
Traveling_required               0.35103    0.29994   1.170              0.24187
NoticePeriod_buyout              0.24619    0.37240   0.661              0.50856
Job_type                         2.18562    0.31841   6.864  0.00000000000668408 ***
Whether_comp_Possess_vehicle     3.60466    0.45451   7.931  0.00000000000000218 ***
Work_location                   -0.01947    0.29791  -0.065              0.94790
Traveltowork_distance_._time    -0.04775    0.20264  -0.236              0.81372
Work_area_living_cost            0.43146    0.26047   1.656              0.09763 .
Public_transport_nearby          3.07805    0.33344   9.231 < 0.0000000000000002 ***
```

```
Current_employment              -1.08451    0.37851  -2.865               0.00417 **
Offer_letter_Processtime.week.   3.29408    0.28909  11.395 < 0.0000000000000002 ***
Bonus                            4.90706    0.50905   9.640 < 0.0000000000000002 ***
Allowance                       -0.34537    0.45873  -0.753               0.45153
Overtime_pay                     2.55933    0.39446   6.488  0.0000000000869730 ***
Emp_shares_scheme               -2.81092    0.47414  -5.928  0.0000000030577627 ***
Performance_incentive            2.45298    0.41452   5.918  0.0000000032645177 ***
Annual_leave                     5.28950    0.92119   5.742  0.0000000093555924 ***
Medical_Family                   5.75821    0.52312  11.008 < 0.0000000000000002 ***
Child_day_care                  -1.46531    0.48699  -3.009               0.00262 **
Nursing_room                    10.72449    0.90730  11.820 < 0.0000000000000002 ***
Parking                          3.47459    0.39378   8.824 < 0.0000000000000002 ***
Comp_size                       -4.79128    0.36828 -13.010 < 0.0000000000000002 ***
Comp_rating                      1.79121    0.20912   8.565 < 0.0000000000000002 ***
Work_life                        1.14837    0.18141   6.330  0.0000000002446575 ***
Stresslevel                     -1.22270    0.16110  -7.590  0.0000000000000320 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 4880.97  on 3526  degrees of freedom
Residual deviance:  862.09  on 3494  degrees of freedom
AIC: 928.09

Number of Fisher Scoring iterations: 9
```

The null deviance value is very high which can say that the model is not so good. But apart from that model, we can have all the variables which are directly contributing to the dependent variable except some like allowance, work location, travel from office time, work area living cost, job type, notice period buyout.

Null Deviance = 2(LL(Saturated Model) - LL(Null Model)) on df = df_Sat - df_Null
Residual Deviance = 2(LL(Saturated Model) - LL(Proposed Model)) df = df_Sat - df_Proposed
The **Saturated Model** is a model that assumes each data point has its own parameters (which means you have n parameters to estimate.)
The **Null Model** assumes the exact "opposite", in that is assumes one parameter for all of the data points, which means you only estimate 1 parameter.
The **Proposed Model** assumes you can explain your data points with p parameters + an intercept term, so you have p+1 parameters.
If your **Null Deviance** is really small, it means that the Null Model explains the data pretty well. Likewise with your **Residual Deviance**.
What does really small mean? If your model is "good" then your **Deviance** is approx Chi^2 with (df_sat - df_model) degrees of freedom.
If you want to compare you Null model with your Proposed model, then you can look at **(Null Deviance - Residual Deviance)** approx Chi^2 with **df Proposed - df Null** = (n-(p+1))-(n-1)=p

The degrees of freedom reported on the Null are always higher than the degrees of freedom reported on the Residual. That is because again, Null Deviance df = Saturated df - Null df = n-1 Residual Deviance df = Saturated df - Proposed df = n-(p+1)