HR Data Analysis

Aksana Sutyrka

2023-10-31

Installing packages

```
install.packages("skimr")
install.packages("ggcorrplot")
library ("ggcorrplot")
library ("tidyverse")
library ("here")
library ("ggplot2")
library ("janitor")
library ("qrng")
library ("RColorBrewer")
library ("dplyr")
library ("skimr")
```

Setting my favourite plot theme

```
theme_set (theme_classic())
```

Importing the dataset

```
hr <- read.csv ("HR-Employee-Attrition.csv")
```

The task is to plot a correlation map for all numeric variables.

·Overtime ·Marital Status ·Job Role ·Gender ·Education Field ·Department ·Business Travel ·Relation between Overtime and Age ·Total Working Years ·Education Level ·Number of Companies Worked ·Distance from Home

Exploring the dataset

```
glimpse (hr)
```

```
## Rows: 1,470
## Columns: 35
## $ Age                      <int> 41, 49, 37, 33, 27, 32, 59, 30, 38, 36, 35, 2…
## $ Attrition                <chr> "Yes", "No", "Yes", "No", "No", "No", "No", "…
## $ BusinessTravel           <chr> "Travel_Rarely", "Travel_Frequently", "Travel…
## $ DailyRate                <int> 1102, 279, 1373, 1392, 591, 1005, 1324, 1358,…
## $ Department               <chr> "Sales", "Research & Development", "Research …
## $ DistanceFromHome         <int> 1, 8, 2, 3, 2, 2, 3, 24, 23, 27, 16, 15, 26, …
## $ Education                <int> 2, 1, 2, 4, 1, 2, 3, 1, 3, 3, 3, 2, 1, 2, 3, …
## $ EducationField           <chr> "Life Sciences", "Life Sciences", "Other", "L…
## $ EmployeeCount            <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …
## $ EmployeeNumber           <int> 1, 2, 4, 5, 7, 8, 10, 11, 12, 13, 14, 15, 16,…
## $ EnvironmentSatisfaction  <int> 2, 3, 4, 4, 1, 4, 3, 4, 4, 3, 1, 4, 1, 2, 3, …
## $ Gender                   <chr> "Female", "Male", "Male", "Female", "Male", "…
## $ HourlyRate               <int> 94, 61, 92, 56, 40, 79, 81, 67, 44, 94, 84, 4…
## $ JobInvolvement           <int> 3, 2, 2, 3, 3, 3, 4, 3, 2, 3, 4, 2, 3, 3, 2, …
## $ JobLevel                 <int> 2, 2, 1, 1, 1, 1, 1, 1, 3, 2, 1, 2, 1, 1, 1, …
## $ JobRole                  <chr> "Sales Executive", "Research Scientist", "Lab…
## $ JobSatisfaction          <int> 4, 2, 3, 3, 2, 4, 1, 3, 3, 3, 2, 3, 3, 4, 3, …
## $ MaritalStatus            <chr> "Single", "Married", "Single", "Married", "Ma…
## $ MonthlyIncome            <int> 5993, 5130, 2090, 2909, 3468, 3068, 2670, 269…
## $ MonthlyRate              <int> 19479, 24907, 2396, 23159, 16632, 11864, 9964…
## $ NumCompaniesWorked       <int> 8, 1, 6, 1, 9, 0, 4, 1, 0, 6, 0, 0, 1, 0, 5, …
## $ Over18                   <chr> "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", …
## $ OverTime                 <chr> "Yes", "No", "Yes", "Yes", "No", "No", "Yes",…
## $ PercentSalaryHike        <int> 11, 23, 15, 11, 12, 13, 20, 22, 21, 13, 13, 1…
## $ PerformanceRating        <int> 3, 4, 3, 3, 3, 3, 4, 4, 4, 3, 3, 3, 3, 3, 3, …
## $ RelationshipSatisfaction <int> 1, 4, 2, 3, 4, 3, 1, 2, 2, 2, 3, 4, 4, 3, 2, …
## $ StandardHours            <int> 80, 80, 80, 80, 80, 80, 80, 80, 80, 80, 80, 8…
## $ StockOptionLevel         <int> 0, 1, 0, 0, 1, 0, 3, 1, 0, 2, 1, 0, 1, 1, 0, …
## $ TotalWorkingYears        <int> 8, 10, 7, 8, 6, 8, 12, 1, 10, 17, 6, 10, 5, 3…
## $ TrainingTimesLastYear    <int> 0, 3, 3, 3, 3, 2, 3, 2, 2, 3, 5, 3, 1, 2, 4, …
## $ WorkLifeBalance          <int> 1, 3, 3, 3, 3, 2, 2, 3, 3, 2, 3, 3, 2, 3, 3, …
## $ YearsAtCompany           <int> 6, 10, 0, 8, 2, 7, 1, 1, 9, 7, 5, 9, 5, 2, 4,…
## $ YearsInCurrentRole       <int> 4, 7, 0, 7, 2, 7, 0, 0, 7, 7, 4, 5, 2, 2, 2, …
## $ YearsSinceLastPromotion  <int> 0, 1, 0, 3, 2, 3, 0, 0, 1, 7, 0, 0, 4, 1, 0, …
## $ YearsWithCurrManager     <int> 5, 7, 0, 0, 2, 6, 0, 0, 8, 7, 3, 8, 3, 2, 3, …
```

Cleaning the dataset

Renaming columns (cleaning names)

```
hr <- clean_names (hr)
```

Creating factors

```
hr <- hr %>%
   mutate_if(sapply(hr, is.character), as.factor)
glimpse (hr)
```

```
## Rows: 1,470
## Columns: 35
## $ age                          <int> 41, 49, 37, 33, 27, 32, 59, 30, 38, 36, 35,…
## $ attrition                    <fct> Yes, No, Yes, No, No, No, No, No, No, No, N…
## $ business_travel              <fct> Travel_Rarely, Travel_Frequently, Travel_Ra…
## $ daily_rate                   <int> 1102, 279, 1373, 1392, 591, 1005, 1324, 135…
## $ department                   <fct> Sales, Research & Development, Research & D…
## $ distance_from_home           <int> 1, 8, 2, 3, 2, 2, 3, 24, 23, 27, 16, 15, 26…
## $ education                    <int> 2, 1, 2, 4, 1, 2, 3, 1, 3, 3, 3, 2, 1, 2, 3…
## $ education_field              <fct> Life Sciences, Life Sciences, Other, Life S…
## $ employee_count               <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
## $ employee_number              <int> 1, 2, 4, 5, 7, 8, 10, 11, 12, 13, 14, 15, 1…
## $ environment_satisfaction     <int> 2, 3, 4, 4, 1, 4, 3, 4, 4, 3, 1, 4, 1, 2, 3…
## $ gender                       <fct> Female, Male, Male, Female, Male, Male, Fem…
## $ hourly_rate                  <int> 94, 61, 92, 56, 40, 79, 81, 67, 44, 94, 84,…
## $ job_involvement              <int> 3, 2, 2, 3, 3, 3, 4, 3, 2, 3, 4, 2, 3, 3, 2…
## $ job_level                    <int> 2, 2, 1, 1, 1, 1, 1, 1, 3, 2, 1, 2, 1, 1, 1…
## $ job_role                     <fct> Sales Executive, Research Scientist, Labora…
## $ job_satisfaction             <int> 4, 2, 3, 3, 2, 4, 1, 3, 3, 3, 2, 3, 3, 4, 3…
## $ marital_status               <fct> Single, Married, Single, Married, Married, …
## $ monthly_income               <int> 5993, 5130, 2090, 2909, 3468, 3068, 2670, 2…
## $ monthly_rate                 <int> 19479, 24907, 2396, 23159, 16632, 11864, 99…
## $ num_companies_worked         <int> 8, 1, 6, 1, 9, 0, 4, 1, 0, 6, 0, 0, 1, 0, 5…
## $ over18                       <fct> Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, Y…
## $ over_time                    <fct> Yes, No, Yes, Yes, No, No, Yes, No, No, No,…
## $ percent_salary_hike          <int> 11, 23, 15, 11, 12, 13, 20, 22, 21, 13, 13,…
## $ performance_rating           <int> 3, 4, 3, 3, 3, 3, 4, 4, 4, 3, 3, 3, 3, 3, 3…
## $ relationship_satisfaction    <int> 1, 4, 2, 3, 4, 3, 1, 2, 2, 2, 3, 4, 4, 3, 2…
## $ standard_hours               <int> 80, 80, 80, 80, 80, 80, 80, 80, 80, 80, 80,…
## $ stock_option_level           <int> 0, 1, 0, 0, 1, 0, 3, 1, 0, 2, 1, 0, 1, 1, 0…
## $ total_working_years          <int> 8, 10, 7, 8, 6, 8, 12, 1, 10, 17, 6, 10, 5,…
## $ training_times_last_year     <int> 0, 3, 3, 3, 3, 2, 3, 2, 2, 3, 5, 3, 1, 2, 4…
## $ work_life_balance            <int> 1, 3, 3, 3, 3, 2, 2, 3, 3, 2, 3, 3, 2, 3, 3…
## $ years_at_company             <int> 6, 10, 0, 8, 2, 7, 1, 1, 9, 7, 5, 9, 5, 2, …
## $ years_in_current_role        <int> 4, 7, 0, 7, 2, 7, 0, 0, 7, 7, 4, 5, 2, 2, 2…
## $ years_since_last_promotion <int> 0, 1, 0, 3, 2, 3, 0, 0, 1, 7, 0, 0, 4, 1, 0…
## $ years_with_curr_manager      <int> 5, 7, 0, 0, 2, 6, 0, 0, 8, 7, 3, 8, 3, 2, 3…
```

Getting rid of unnecessary columns

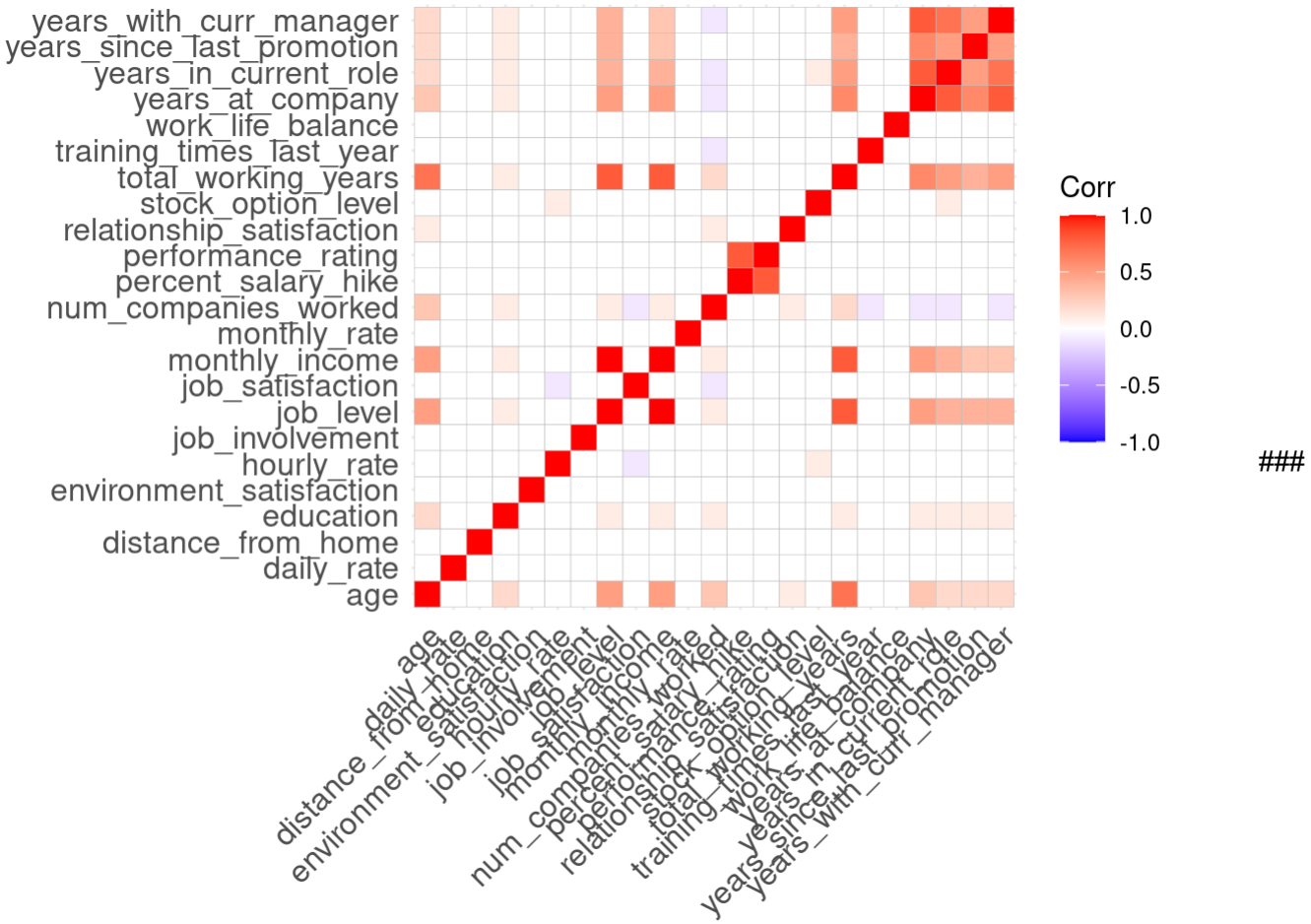
```
hr_corr <- hr %>% select(-c(attrition, business_travel, department, education_field, employee_count, ge
nder, job_role, marital_status, over18, over_time,standard_hours, employee_number))

glimpse (hr_corr)
```

```
## Rows: 1,470
## Columns: 23
## $ age                          <int> 41, 49, 37, 33, 27, 32, 59, 30, 38, 36, 35,…
## $ daily_rate                   <int> 1102, 279, 1373, 1392, 591, 1005, 1324, 135…
## $ distance_from_home           <int> 1, 8, 2, 3, 2, 2, 3, 24, 23, 27, 16, 15, 26…
## $ education                    <int> 2, 1, 2, 4, 1, 2, 3, 1, 3, 3, 3, 2, 1, 2, 3…
## $ environment_satisfaction     <int> 2, 3, 4, 4, 1, 4, 3, 4, 4, 3, 1, 4, 1, 2, 3…
## $ hourly_rate                  <int> 94, 61, 92, 56, 40, 79, 81, 67, 44, 94, 84,…
## $ job_involvement              <int> 3, 2, 2, 3, 3, 3, 4, 3, 2, 3, 4, 2, 3, 3, 2…
## $ job_level                    <int> 2, 2, 1, 1, 1, 1, 1, 1, 3, 2, 1, 2, 1, 1, 1…
## $ job_satisfaction             <int> 4, 2, 3, 3, 2, 4, 1, 3, 3, 3, 2, 3, 3, 4, 3…
## $ monthly_income               <int> 5993, 5130, 2090, 2909, 3468, 3068, 2670, 2…
## $ monthly_rate                 <int> 19479, 24907, 2396, 23159, 16632, 11864, 99…
## $ num_companies_worked         <int> 8, 1, 6, 1, 9, 0, 4, 1, 0, 6, 0, 0, 1, 0, 5…
## $ percent_salary_hike          <int> 11, 23, 15, 11, 12, 13, 20, 22, 21, 13, 13,…
## $ performance_rating           <int> 3, 4, 3, 3, 3, 3, 4, 4, 4, 3, 3, 3, 3, 3, 3…
## $ relationship_satisfaction    <int> 1, 4, 2, 3, 4, 3, 1, 2, 2, 2, 3, 4, 4, 3, 2…
## $ stock_option_level           <int> 0, 1, 0, 0, 1, 0, 3, 1, 0, 2, 1, 0, 1, 1, 0…
## $ total_working_years          <int> 8, 10, 7, 8, 6, 8, 12, 1, 10, 17, 6, 10, 5,…
## $ training_times_last_year     <int> 0, 3, 3, 3, 3, 2, 3, 2, 2, 3, 5, 3, 1, 2, 4…
## $ work_life_balance            <int> 1, 3, 3, 3, 3, 2, 2, 3, 3, 2, 3, 3, 2, 3, 3…
## $ years_at_company             <int> 6, 10, 0, 8, 2, 7, 1, 1, 9, 7, 5, 9, 5, 2, …
## $ years_in_current_role        <int> 4, 7, 0, 7, 2, 7, 0, 0, 7, 7, 4, 5, 2, 2, 2…
## $ years_since_last_promotion <int> 0, 1, 0, 3, 2, 3, 0, 0, 1, 7, 0, 0, 4, 1, 0…
## $ years_with_curr_manager      <int> 5, 7, 0, 0, 2, 6, 0, 0, 8, 7, 3, 8, 3, 2, 3…
```

Creating a correlation plot

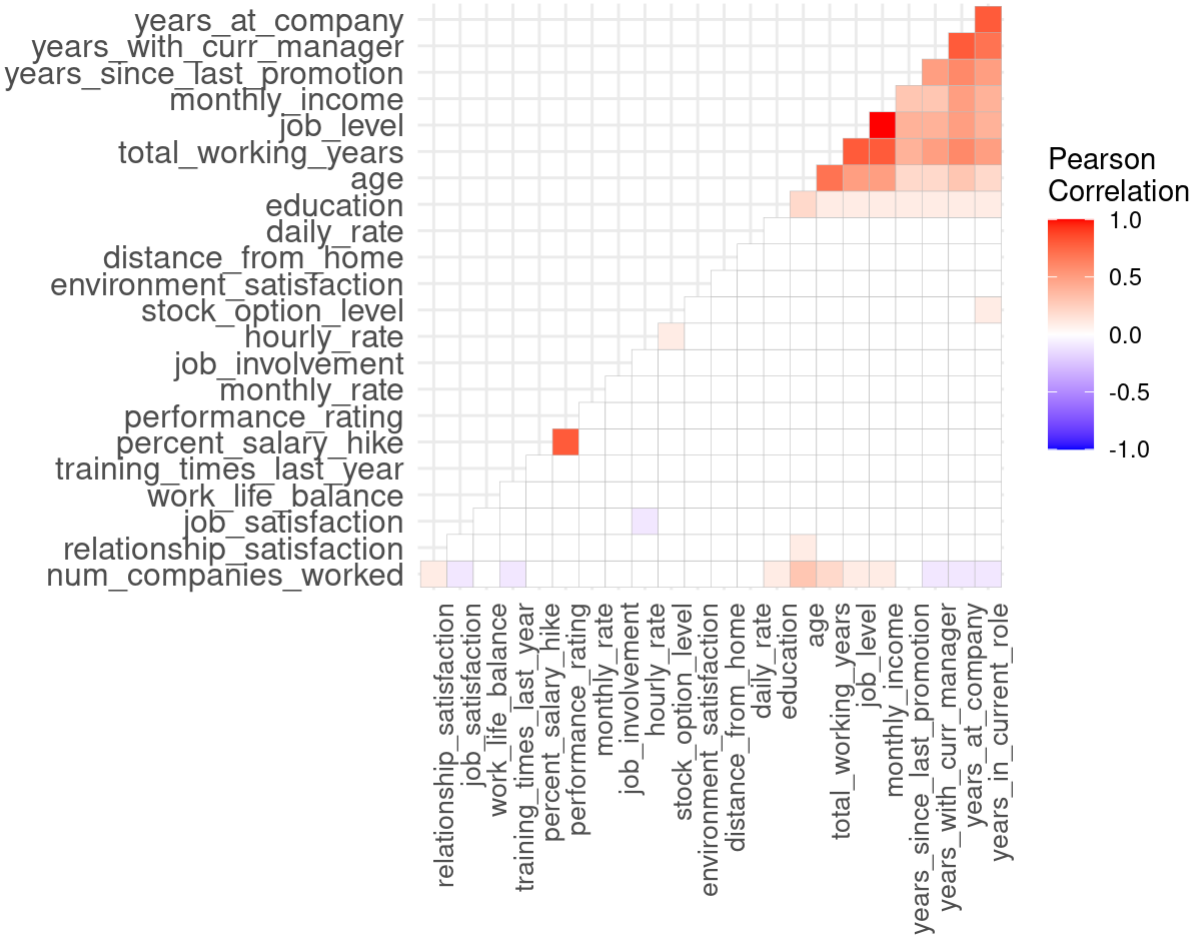
```
correlation_matrix <- round (cor(hr_corr),1)
ggcorrplot(correlation_matrix, method ="square")
```



Visualizing triangular upper plot

```
p.mat <- cor_pmat(hr_corr)
ggcorrplot(correlation_matrix,
           p.mat = p.mat,
  hc.order = TRUE,
  type = "lower",
  insig = "blank")+
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
   midpoint = 0, limit = c(-1,1), space = "Lab",
    name="Pearson\nCorrelation") +
   theme(axis.text.x = element_text(angle = 90, vjust = 1,
    size = 10, hjust = 1))
```

```
## Scale for fill is already present.
## Adding another scale for fill, which will replace the existing scale.
```



Deleting unnecessary columns (with correlation coefficient <0.5) from this correlation_matrix to make it neater

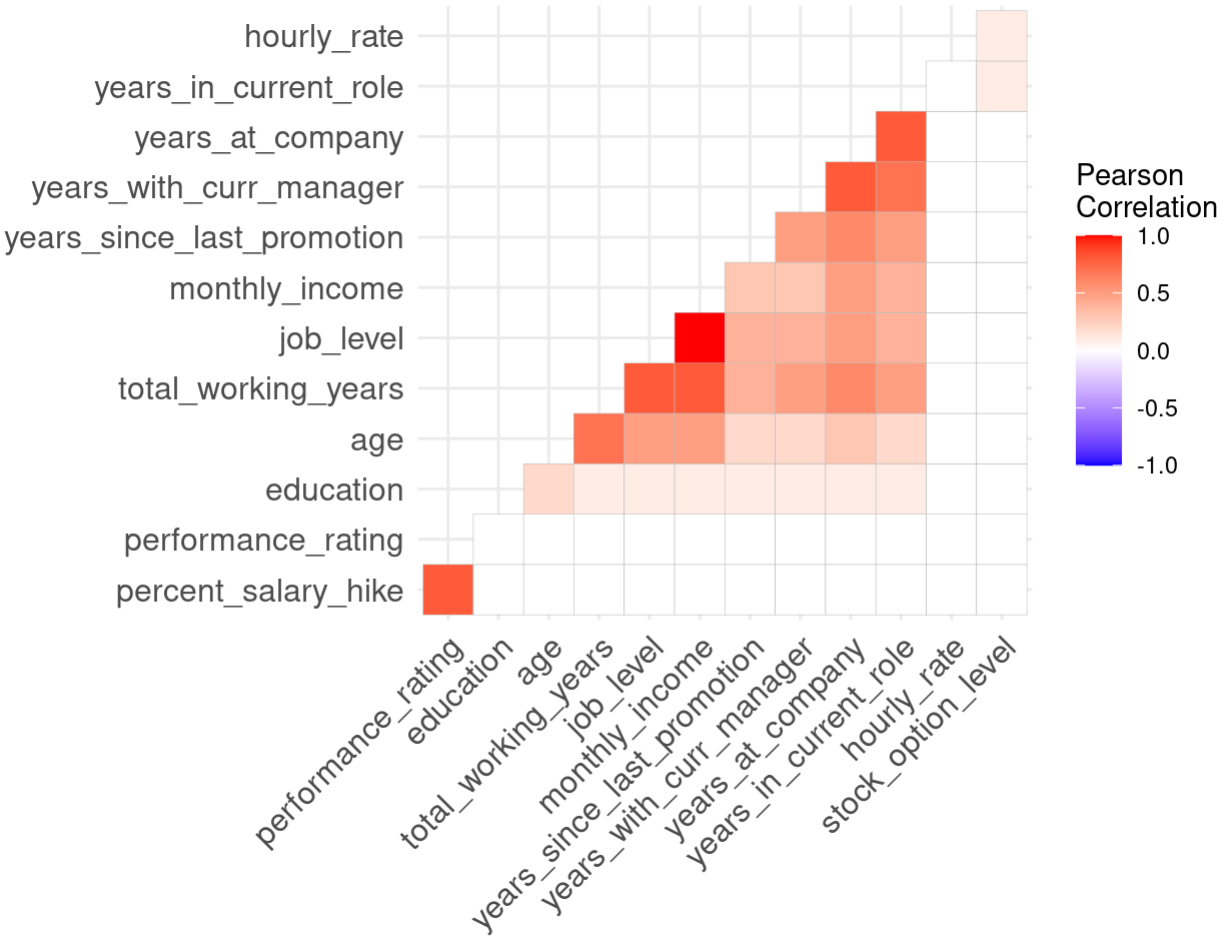
```
hr_corr <- hr_corr %>% select(-c(work_life_balance,environment_satisfaction,distance_from_home,
                                 job_involvement, daily_rate, monthly_rate, training_times_last_year, j
ob_satisfaction, relationship_satisfaction, num_companies_worked ))
glimpse (hr_corr)
```

```
## Rows: 1,470
## Columns: 13
## $ age                        <int> 41, 49, 37, 33, 27, 32, 59, 30, 38, 36, 35,…
## $ education                  <int> 2, 1, 2, 4, 1, 2, 3, 1, 3, 3, 3, 2, 1, 2, 3…
## $ hourly_rate                <int> 94, 61, 92, 56, 40, 79, 81, 67, 44, 94, 84,…
## $ job_level                  <int> 2, 2, 1, 1, 1, 1, 1, 1, 3, 2, 1, 2, 1, 1, 1…
## $ monthly_income             <int> 5993, 5130, 2090, 2909, 3468, 3068, 2670, 2…
## $ percent_salary_hike        <int> 11, 23, 15, 11, 12, 13, 20, 22, 21, 13, 13,…
## $ performance_rating         <int> 3, 4, 3, 3, 3, 3, 4, 4, 4, 3, 3, 3, 3, 3, 3…
## $ stock_option_level         <int> 0, 1, 0, 0, 1, 0, 3, 1, 0, 2, 1, 0, 1, 1, 0…
## $ total_working_years        <int> 8, 10, 7, 8, 6, 8, 12, 1, 10, 17, 6, 10, 5,…
## $ years_at_company           <int> 6, 10, 0, 8, 2, 7, 1, 1, 9, 7, 5, 9, 5, 2, …
## $ years_in_current_role      <int> 4, 7, 0, 7, 2, 7, 0, 0, 7, 7, 4, 5, 2, 2, 2…
## $ years_since_last_promotion <int> 0, 1, 0, 3, 2, 3, 0, 0, 1, 7, 0, 0, 4, 1, 0…
## $ years_with_curr_manager    <int> 5, 7, 0, 0, 2, 6, 0, 0, 8, 7, 3, 8, 3, 2, 3…
```

```
correlation_matrix <- round (cor(hr_corr),1)
ggcorrplot(correlation_matrix, method ="square")
```


```
p.mat <- cor_pmat(hr_corr)
ggcorrplot(correlation_matrix,
           p.mat = p.mat,
  hc.order = TRUE,
  type = "lower",
  insig = "blank")+
scale_fill_gradient2(low = "blue", high = "red", mid = "white",
   midpoint = 0, limit = c(-1,1), space = "Lab",
    name="Pearson\nCorrelation")
```

```
## Scale for fill is already present.
## Adding another scale for fill, which will replace the existing scale.
```



We can see strong positive correlation between job level and monthly income, total working years and monthly income and job level, age and total working years, which is quite expected.There's also strong correlation between percent salary hike and performance rating which means that for better performance you are better paid.