ertugrulturkseven_Project_1

2024-08-03

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

\

```
#file.choose()

project_1 <- read.csv("/Users/ertuboston/Documents/Data_Science_Merrimack/DSE5002/PROJECT_1/r project da
```

```
library(readr)
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(stringr)
library(scales)
```

```
##
## Attaching package: 'scales'

## The following object is masked from 'package:readr':
##
##     col_factor
```

```
 # ------------------------CREATING SUBSETS FOR DATA SCIENTIST---------------------- #

### Created subset based on the job title.
subset_data_science <- subset(project_1, (project_1$job_title == "Data Scientist" & project_1$employmen
names(subset_data_science)
```

```
##  [1] "X"                   "work_year"           "experience_level"
##  [4] "employment_type"     "job_title"           "salary"
##  [7] "salary_currency"     "salary_in_usd"       "employee_residence"
## [10] "remote_ratio"        "company_location"    "company_size"
```

```
### replaced the employee residence (the ones are not US with OffShore)
subset_data_science<- subset_data_science %>%
  mutate(employee_residence= if_else(employee_residence != "US", "OffShore", employee_residence ))


### created subset df based on the company size
small_comp_df <- subset (subset_data_science, (subset_data_science$company_size == "S"))

medium_comp_df <- subset (subset_data_science, (subset_data_science$company_size == "M"))

large_comp_df <- subset (subset_data_science, (subset_data_science$company_size == "L"))
```

```
# ------------------------CREATE SMALL COMPANIES DF + GRAPHING ---------------------- #


### Created av_sall_small DF grouped by for the specific columns for Small size comp.,
### and created average, max and min salary columns

av_sall_small <- small_comp_df %>%
  group_by(employee_residence,work_year, experience_level,company_size) %>%
  summarise(average_salary = mean(salary_in_usd),
            max_salary = max(salary_in_usd),
            min_salary = min(salary_in_usd))
```

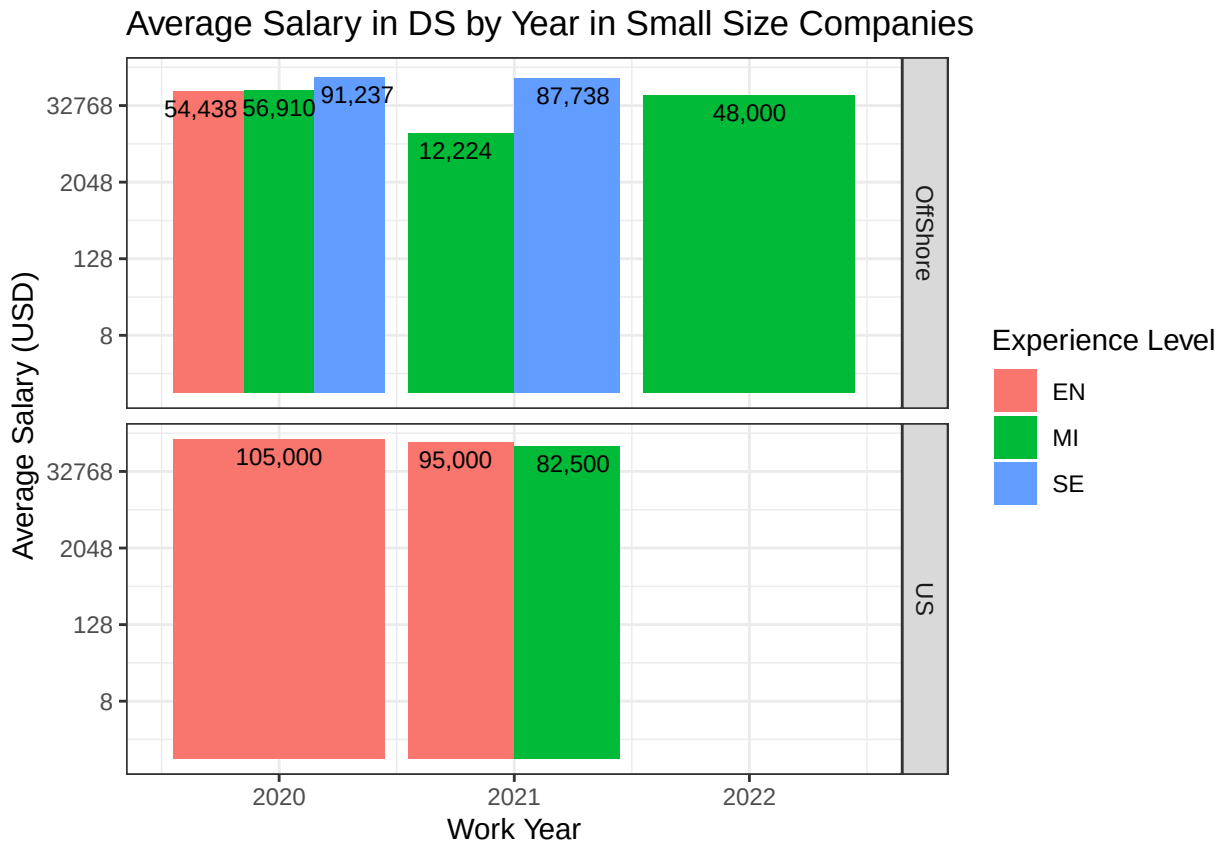
```
## `summarise()` has grouped output by 'employee_residence', 'work_year',
## 'experience_level'. You can override using the `.groups` argument.
```

```
av_sall_small$salary_range <- paste("$",av_sall_small$min_salary, "-", "$",av_sall_small$max_salary)

# In this code I wanted to see the average salaries without grouping them by year.
salary_average_noyear<- small_comp_df %>% group_by(experience_level, employee_residence) %>%
  summarise(average = mean(salary_in_usd),
            min_sal = min(salary_in_usd),
            max_sal = max(salary_in_usd))
```

```
## `summarise()` has grouped output by 'experience_level'. You can override using
## the `.groups` argument.
```

```
# I wanted to work on excel(google Sheet) to see what I can do in that table.
write.csv(salary_average_noyear, "average salary between 2020-2022.csv",row.names = FALSE)
#write.csv(av_sall_small, "salary_table.csv", row.names = FALSE)

### Here, I am graphing the small companies average salary by residence, experience and year
### I added experience level in to the graph by creating facet_grid for experience level.
graph_small <- ggplot(av_sall_small, aes(x=work_year,y = average_salary, fill = experience_level)) +
  geom_col(position = "dodge") + facet_grid("employee_residence") +
  scale_y_continuous(trans = 'log2') +
```

```
  geom_text(aes(label=label_comma() (round(average_salary))), vjust = 1.5, size =3,
            color = "black", position = position_dodge(width = 1)) +
  labs(title = "Average Salary in DS by Year in Small Size Companies",
       x = "Work Year",
       y = "Average Salary (USD)",
       fill = "Experience Level") +
  theme_bw()


print(graph_small)
```

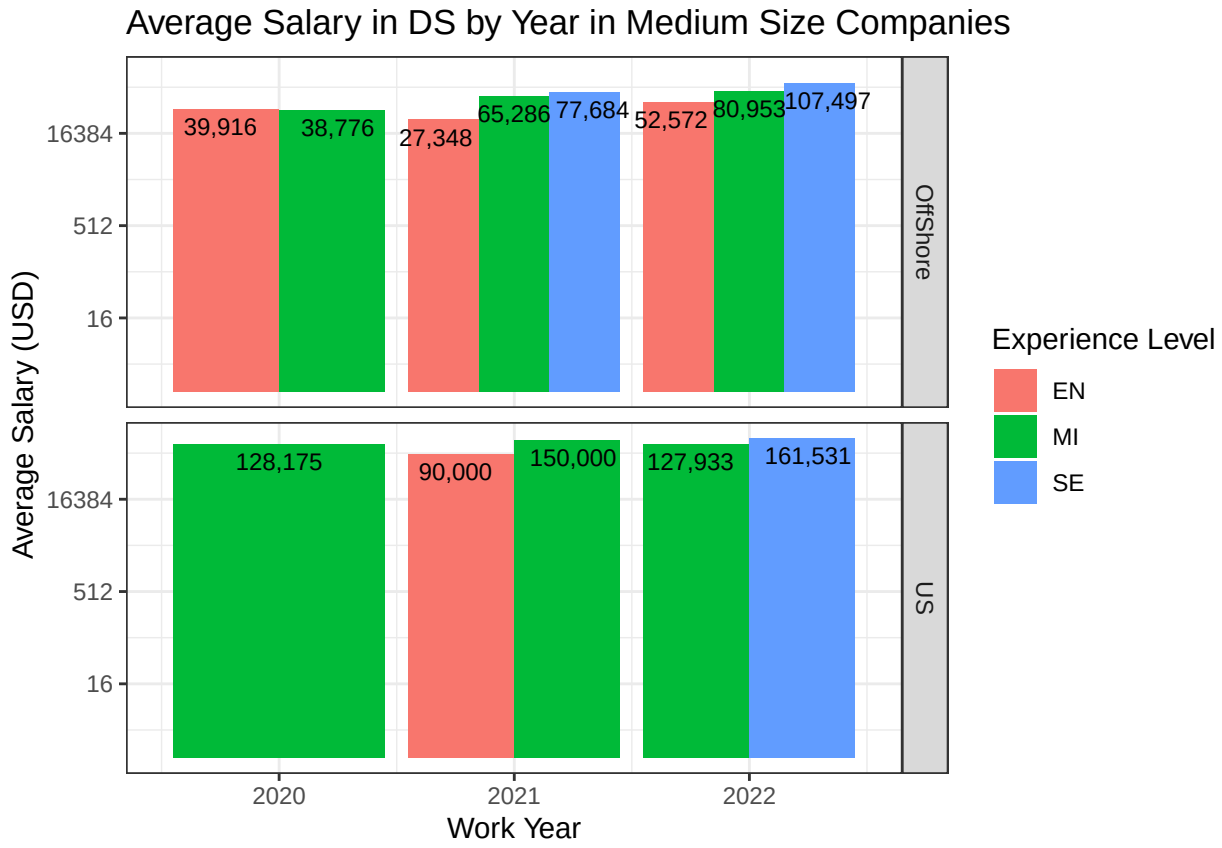



```
# ------------------------CREATE MEDIUM COMPANIES DF + GRAPHING ---------------------- #


### Creating average salary for medium size companies
av_sall_medium <- medium_comp_df %>%
  group_by(employee_residence, work_year, experience_level,company_size) %>%
  summarise(average_salary = mean(salary_in_usd))
```

```
## `summarise()` has grouped output by 'employee_residence', 'work_year',
## 'experience_level'. You can override using the `.groups` argument.
```

```
### here, I am graphing the medium companies average salary by residence, experience and year
### I added experience level in to the graph by creating facet_grid for experience level.
```

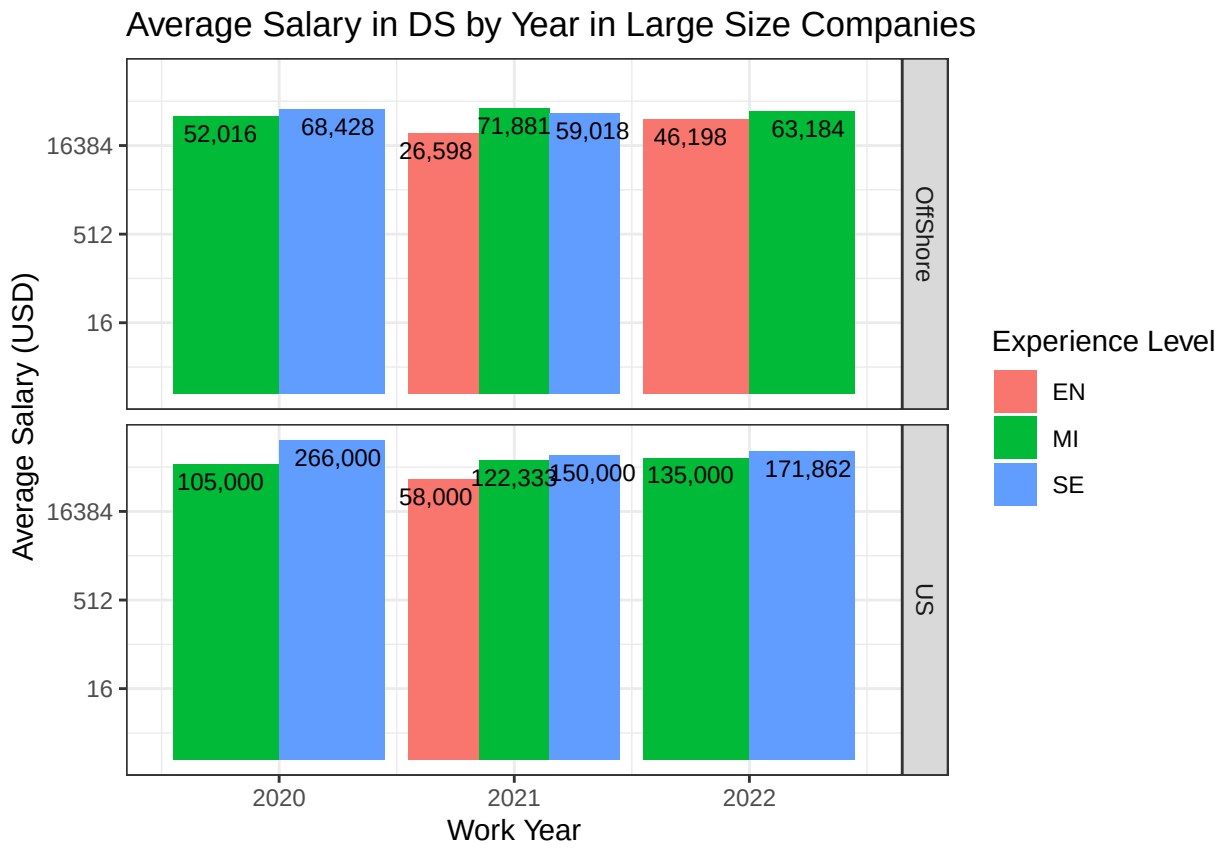
```
graph_medium <- ggplot(av_sall_medium, aes(x=work_year,y = average_salary, fill = experience_level))+
  # color = as.factor(work_year))) +
  geom_col(position = "dodge") + facet_grid("employee_residence") +
  scale_y_continuous(trans = 'log2') +
  geom_text(aes(label=label_comma() (round(average_salary))), vjust = 1.5, size =3,
            color = "black", position = position_dodge(width = 1)) +
  labs(title = "Average Salary in DS by Year in Medium Size Companies",
       x = "Work Year",
       y = "Average Salary (USD)",
       fill = "Experience Level") +
  theme_bw()

print(graph_medium)
```




```
# ------------------------CREATE LARGE COMPANIES DF + GRAPHING ---------------------- #

### Created av_sall_small DF grouped by for the specific columns for Large size companies,
### and created average, max and min salary columns

av_sall_large <- large_comp_df %>%
  group_by(employee_residence, work_year,experience_level,company_size) %>%
  summarise(average_salary = mean(salary_in_usd))
```

```
## `summarise()` has grouped output by 'employee_residence', 'work_year',
## 'experience_level'. You can override using the `.groups` argument.
```

```r
### Here, I am graphing the large companies average salary by residence, experience and year
### I added experience level in to the graph by creating facet_grid for experience level.
graph_large <- ggplot(av_sall_large, aes(x=work_year,y = average_salary, fill = experience_level))+
  # color = as.factor(work_year))) +
  geom_col(position = "dodge") + facet_grid("employee_residence") +
  scale_y_continuous(trans = 'log2') +
  geom_text(aes(label=label_comma()(round(average_salary))), vjust = 1.5, size =3,
            color = "black", position = position_dodge(width = 1)) +
  labs(title = "Average Salary in DS by Year in Large Size Companies",
       x = "Work Year",
       y = "Average Salary (USD)",
       fill = "Experience Level") +
  theme_bw()

print(graph_large)
```

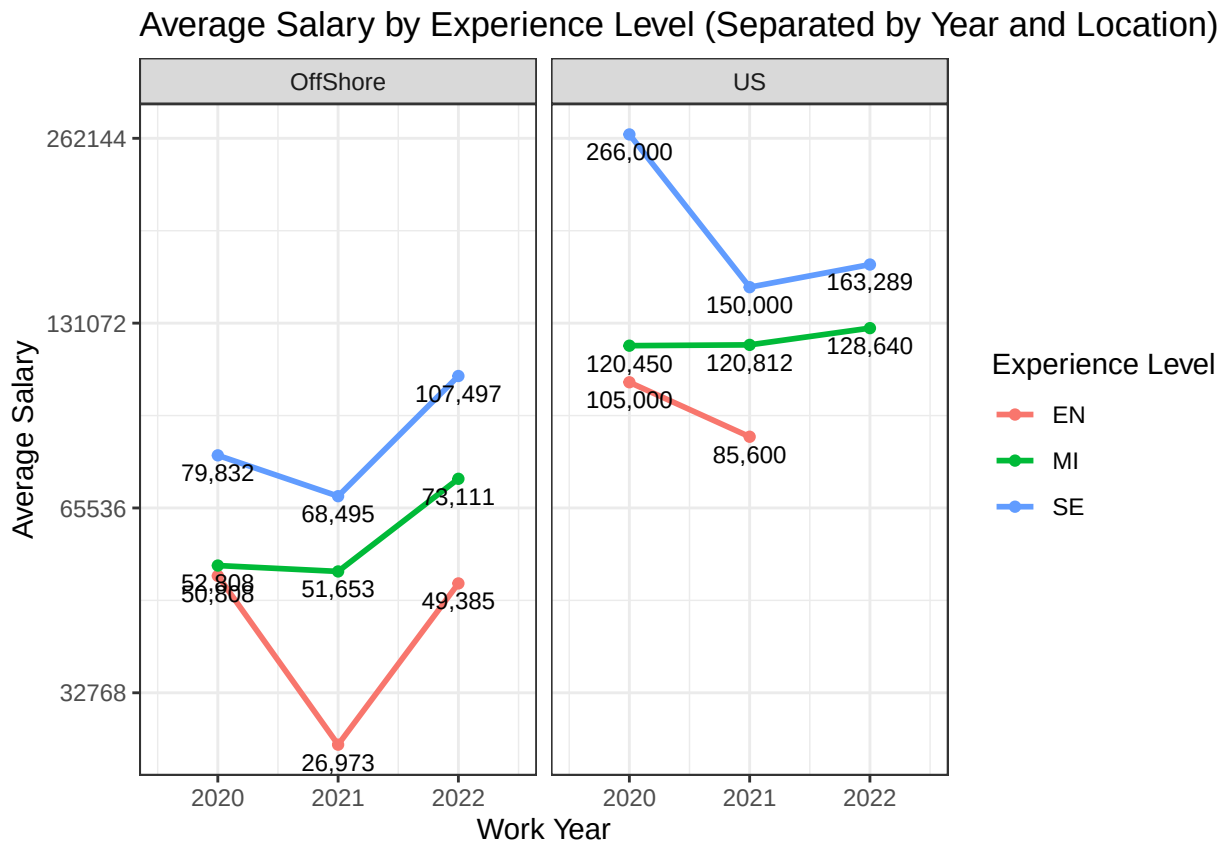



```r
# ------------------------CREATE LINE GRAPH COMPANIES DF + GRAPHING ---------------------- #

### Here, I am creating a line graph to show the difference in average salary in years.
### subset_data_science has only data science with FT position, and it will also have
### employee residence in US and OffShore. I did not group them by company size.
### This is just to show the DS average salaries in the market
### without separating them by company size.

av_sall_DS <- subset_data_science %>%
```

```
  group_by(work_year,employee_residence, experience_level) %>%
  summarise(av_sall = mean(salary_in_usd))
```

```
## `summarise()` has grouped output by 'work_year', 'employee_residence'. You can
## override using the `.groups` argument.
```

```
### this is the graph where I can see the increases and decreases of the average
### salary in offshore and US by the year.

experience_graph <- ggplot(av_sall_DS, aes(x=work_year, y=av_sall, color = experience_level)) +
  scale_y_continuous(trans ='log2') +
  geom_line(linewidth =1) + facet_wrap("employee_residence") +
  geom_point() +
  geom_text(aes(label=label_comma() (round(av_sall))), vjust = 1.5, size =3,
            color = "black", position = position_dodge(width = 1)) +
  labs(title = "Average Salary by Experience Level (Separated by Year and Location)",
       x = "Work Year",
       y = "Average Salary",
       color ="Experience Level",
       size = 4) +
  theme_bw()

print(experience_graph)
```



```
  #geom_point (size =3)
```

```
### I created another dataframe which is grouped by work year and employee residence only
### to see the salary range overall in years without considering the employee levels
### and the company sizes. I write the dataframe to csv file to work on the excel(google sheet).

salary_range <- subset_data_science %>%
  group_by(work_year, employee_residence) %>%
  summarise(min_salary = min(salary_in_usd), max_salary = max(salary_in_usd))
```

```
## `summarise()` has grouped output by 'work_year'. You can override using the
## `.groups` argument.
```

```
write.csv(salary_range,"salary_range.csv",row.names=FALSE)
```