# Data Analysis

Preksha Patel

# Research Question/ Problem Definition

What is the impact of various factors, such as employees' level of education, gender, age, and years of experience, on salary?

1) How does work experience relate to salary?
2) How does education level affect salary?
3) How does gender relate to salary?
4) How does age affect salary?

# Hypothesis:

- **Level of Education:** Employees with higher levels of education, such as advanced degrees or specialized certifications, are likely to earn higher salaries compared to those with lower levels of education.

- **Gender:** We anticipate that there might be a gender pay gap, with male employees tending to have higher salaries on average compared to female employees, even when controlling for other factors.

- **Age:** Younger employees may have lower salaries due to limited experience, while more experienced employees may demand higher salaries due to their expertise and seniority.

- **Years of Experience:** It is expected that employees with more years of experience will generally receive higher salaries, as their accumulated knowledge and skills are likely to be valued by employers.

# Hypothesis (continued)

- **Level of Education:** The null hypothesis is that there is no significant difference in salaries between the populations. The alternative hypothesis is that there is a difference in the mean salaries for each population.

- **Gender:** The null hypothesis is that there is no significant difference in salaries between the populations (focusing on male and female). The alternative hypothesis is that there is a difference in the mean salaries for each population.

- **Age:** The null hypothesis is that there is no significant difference in salaries between the populations (different age groups). The alternative hypothesis is that there is a difference in the mean salaries for each population.

- **Years of Experience:** The null hypothesis is that there is no significant difference in salaries between the populations (varying years of experience). The alternative hypothesis is that there is a difference in the mean salaries for each population.

# Dataset

This dataset portrayed the salary data of various individuals as well as other data about them including age, gender (female, male, or other), education level (from no school to a PhD), job title, and salary.

This is also our cleaned data set which includes no null or “NA” values. We also added two other columns including “Age_Group” and “Exp_Group” for us to use in analysis and testing.

| | Age | Gender | Education.Level | Job.Title | Years.of.Experience | Salary | Age_Group | Exp_Group |
|---|---|---|---|---|---|---|---|---|
| 1 | 32 | Male | Bachelor's | Software Engineer | 5 | 90000 | 30s | Beginner |
| 2 | 28 | Female | Master's | Data Analyst | 3 | 65000 | 20s | Beginner |
| 3 | 45 | Male | PhD | Senior Manager | 15 | 150000 | 40s | Professional |
| 4 | 36 | Female | Bachelor's | Sales Associate | 7 | 60000 | 30s | Professional |
| 5 | 52 | Male | Master's | Director | 20 | 200000 | 50s | Expert |
| 6 | 29 | Male | Bachelor's | Marketing Analyst | 2 | 55000 | 20s | Beginner |
| 7 | 42 | Female | Master's | Product Manager | 12 | 120000 | 40s | Professional |
| 8 | 31 | Male | Bachelor's | Sales Manager | 4 | 80000 | 30s | Beginner |
| 9 | 26 | Female | Bachelor's | Marketing Coordinator | 1 | 45000 | 20s | Beginner |
| 10 | 38 | Male | PhD | Senior Scientist | 10 | 110000 | 30s | Professional |
| 11 | 29 | Male | Master's | Software Developer | 3 | 75000 | 20s | Beginner |
| 12 | 48 | Female | Bachelor's | HR Manager | 18 | 140000 | 40s | Expert |
| 13 | 35 | Male | Bachelor's | Financial Analyst | 6 | 65000 | 30s | Professional |
| 14 | 40 | Female | Master's | Project Manager | 14 | 130000 | 30s | Professional |
| 15 | 27 | Male | Bachelor's | Customer Service Rep | 2 | 40000 | 20s | Beginner |
| 16 | 44 | Male | Bachelor's | Operations Manager | 16 | 125000 | 40s | Expert |
| 17 | 33 | Female | Master's | Marketing Manager | 7 | 90000 | 30s | Professional |
| 18 | 39 | Male | PhD | Senior Engineer | 12 | 115000 | 30s | Professional |
| 19 | 25 | Female | Bachelor's | Data Entry Clerk | 0 | 35000 | 20s | Beginner |
| 20 | 51 | Male | Bachelor's | Sales Director | 22 | 180000 | 50s | Expert |
| 21 | 34 | Female | Master's | Business Analyst | 5 | 80000 | 30s | Beginner |
| 22 | 47 | Male | Master's | VP of Operations | 19 | 190000 | 40s | Expert |
| 23 | 30 | Male | Bachelor's | IT Support | 2 | 50000 | 20s | Beginner |
| 24 | 36 | Female | Bachelor's | Recruiter | 9 | 60000 | 30s | Professional |
| 25 | 41 | Male | Master's | Financial Manager | 13 | 140000 | 40s | Professional |

# Descriptive Statistics

```
> ##Descriptive statistics
> salaryCSV_clean %>%
+   group_by(Gender) %>%
+   get_summary_stats(Salary, type="mean_sd")
# A tibble: 3 x 5
  Gender variable     n    mean     sd
  <fct>  <fct>    <dbl>   <dbl>  <dbl>
1 Female Salary    3011 107960. 52668.
2 Male   Salary    3670 121456. 52030.
3 Other  Salary      14 125870. 44242.
> ##The gender with the highest average salary is Male.
>
> salaryCSV_clean %>%
+   group_by(Education.Level) %>%
+   get_summary_stats(Salary, type="mean_sd")
# A tibble: 5 x 5
  Education.Level variable     n    mean     sd
  <fct>           <fct>    <dbl>   <dbl>  <dbl>
1 Bachelor's      Salary    3018  95177. 44013.
2 High School     Salary     448  36707. 22549.
3 Master's        Salary    1860 130112. 40641.
4 No Experience   Salary       1 100000     NA
5 PhD             Salary    1368 165772. 34061.
> ##The education level with the highest average salary is the PhD level.
>
```

```
> salaryCSV_clean %>%
+   group_by(Age_Group) %>%
+   get_summary_stats(Salary, type="mean_sd")
# A tibble: 5 x 5
  Age_Group variable     n    mean     sd
  <chr>     <fct>    <dbl>   <dbl>  <dbl>
1 20s       Salary    2890  76690. 39126.
2 30s       Salary    2457 129252. 40366.
3 40s       Salary    1159 170070. 25682.
4 50s       Salary     182 191521. 16890.
5 60s       Salary       7 200000      0
> ##The age group with the highest average salary is individuals in their 60s.
>
> salaryCSV_clean %>%
+   group_by(Exp_Group) %>%
+   get_summary_stats(Salary, type="mean_sd")
# A tibble: 4 x 5
  Exp_Group    variable     n    mean     sd
  <chr>        <fct>    <dbl>   <dbl>  <dbl>
1 Beginner     Salary    2802  69153. 32313.
2 Expert       Salary     825 184084. 21194.
3 Professional Salary    2993 137998. 34300.
4 Professor    Salary      75 185481. 13749.
> ##The experience level with the highest average salary is the Professor.
>
```

# Descriptive Statistics (continued)

- We decided to use descriptive statistics and see what the average salary was based on each category.
- For example, in the age category, we saw that as age increased, salary did as well (possibly due to more experience and more skills).
- Similarly, we did descriptive statistics with the gender and salary variable and found that men make the most on average.
- For education, the more education an individual had (on average), the higher their salary was.

# Analysis Methods

- Descriptive statistics - to compare mean salaries
- Bar charts and scatter plots - to visualize the data and see relationships between variables
- Significance testing:
  - Age: One-Way ANOVA
  - Gender: Independent T-test
  - Education Levels: One-Way ANOVA
  - Experience Levels: One-way Anova

# Analysis Methods

```
> #time to run the test since asumptions are met.
> ttest1 <- subset_data %>%
+   t_test(Salary ~ Gender) %>%
+   add_significance()
> ttest1
# A tibble: 3 x 10
  .y.    group1 group2    n1    n2 statistic     df        p    p.adj p.adj.signif
  <chr>  <chr>  <chr>  <int> <int>     <dbl>  <dbl>    <dbl>    <dbl> <chr>
1 Salary Female Male    1687  2299    -8.92  3592.  7.59e-19 2.28e-18 ****
2 Salary Female Other   1687    14    -0.613   13.3 5.5 e- 1 1   e+ 0 ns
3 Salary Male   Other   2299    14     0.582   13.2 5.7 e- 1 1   e+ 0 ns
>
> #acording to the ttest there is a high significant between male and female
>
```

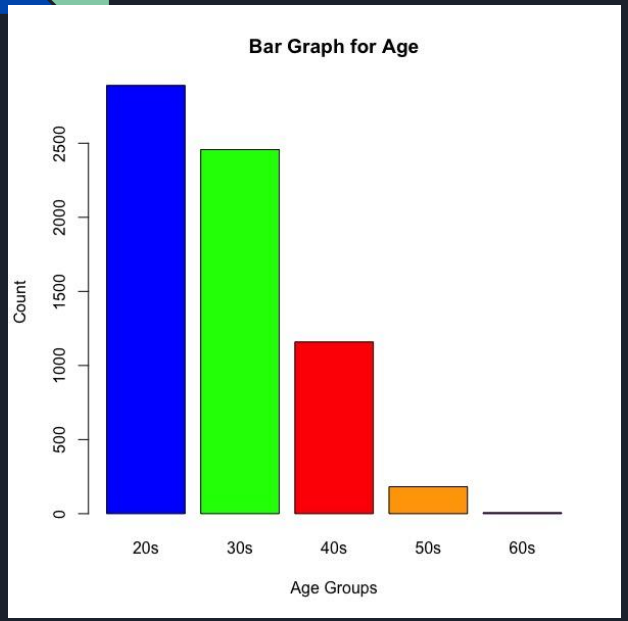
```
> pg.aov <- subset_data %>% anova_test(Salary.L.Log ~ Education.Level)
> # or
> pg.aov
ANOVA Table (type II tests)

          Effect DFn  DFd       F p p<.05   ges
1 Education.Level   4 3995 577.541 0     * 0.366
> pg.pwc <- subset_data %>% tukey_hsd(Salary.L.Log ~ Education.Level)
> pg.pwc
# A tibble: 10 x 9
   term            group1        group2        null.value estimate conf.low conf.high       p.adj p.adj.signif
 * <chr>           <chr>         <chr>              <dbl>    <dbl>    <dbl>     <dbl>       <dbl> <chr>
 1 Education.Level Bachelor's    High School            0  -0.811    -0.907    -0.715    2.49e-8  ****
 2 Education.Level Bachelor's    Master's               0   0.264     0.226     0.303    2.49e-8  ****
 3 Education.Level Bachelor's    No Experience          0   0.0477   -0.995     1.09     1   e+0  ns
 4 Education.Level Bachelor's    PhD                    0   0.599     0.555     0.643    2.49e-8  ****
 5 Education.Level High School   Master's               0   1.08      0.978     1.17     2.49e-8  ****
 6 Education.Level High School   No Experience          0   0.859    -0.188     1.91     1.65e-1  ns
 7 Education.Level High School   PhD                    0   1.41      1.31      1.51     2.49e-8  ****
 8 Education.Level Master's      No Experience          0  -0.217    -1.26      0.827    9.8 e-1  ns
 9 Education.Level Master's      PhD                    0   0.335     0.288     0.381    2.49e-8  ****
10 Education.Level No Experience PhD                    0   0.551    -0.492     1.59     6.01e-1  ns
>
```

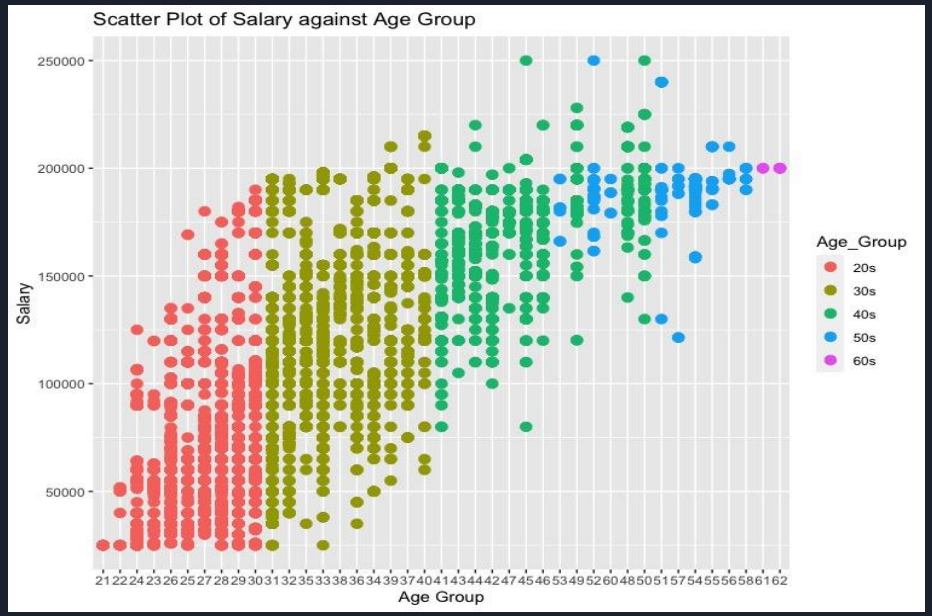
```
> shapiro_test(residuals(Imodel_Exp))
# A tibble: 1 x 3
  variable               statistic  p.value
  <chr>                      <dbl>    <dbl>
1 residuals(Imodel_Exp)     0.992 4.02e-14
>
> subset_data %>%
+   levene_test(Salary ~ Exp_Group)
# A tibble: 1 x 4
    df1   df2 statistic        p
  <int> <int>     <dbl>    <dbl>
1     3  3996      115. 1.53e-71
Warning message:
In leveneTest.default(y = y, group = group, ...) : group coerced to factor.
>
> # transform the data to logarithm and then run the anova_test
> subset_data$Salary.L.Log <- log(subset_data$Salary)
>
>
> pg.aov <- subset_data %>% anova_test(Salary.L.Log ~ Exp_Group)
> # or
> pg.aov
ANOVA Table (type II tests)

     Effect DFn  DFd        F p p<.05   ges
1 Exp_Group   3 3996 1463.002 0     * 0.523
> pg.pwc <- subset_data %>% tukey_hsd(Salary.L.Log ~ Exp_Group)
> pg.pwc
# A tibble: 6 x 9
  term      group1       group2       null.value estimate conf.low conf.high       p.adj p.adj.signif
* <chr>     <chr>        <chr>             <dbl>    <dbl>    <dbl>     <dbl>       <dbl> <chr>
1 Exp_Group Beginner     Expert                0   0.875    0.834     0.916  0.0000000249 ****
2 Exp_Group Beginner     Professional          0   0.621    0.592     0.651  0.0000000249 ****
3 Exp_Group Beginner     Professor             0   0.905    0.805     1.01   0.0000000249 ****
4 Exp_Group Expert       Professional          0  -0.254   -0.294    -0.213  0.0000000249 ****
5 Exp_Group Expert       Professor             0   0.0302  -0.0741    0.135  0.879        ns
6 Exp_Group Professional Professor             0   0.284    0.184     0.384  0.0000000249 ****
>
>
>
```

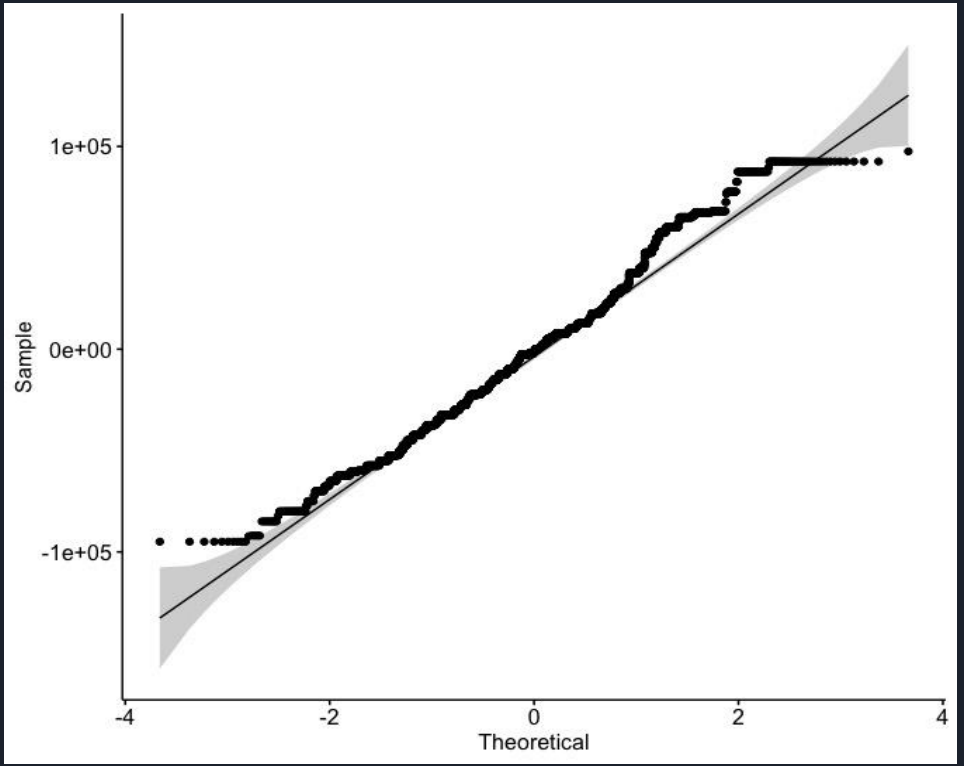
# Graphs (Age)




This graph shows the frequency of each age group within the data so we can better understand how the data is distributed in terms of age.



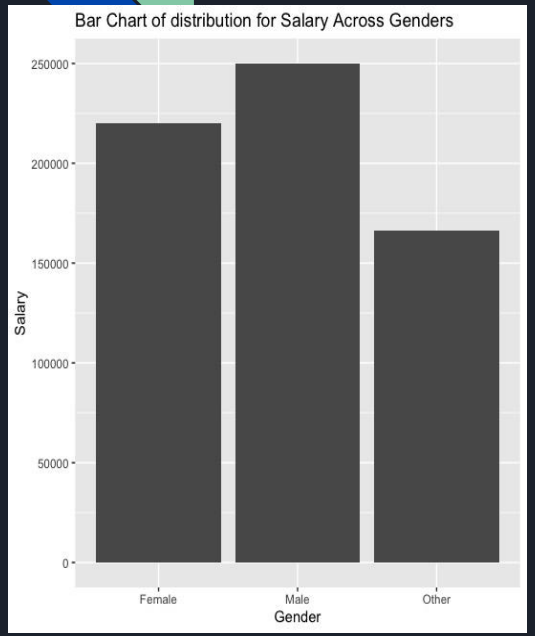

For this graph, we hope to see how the data is distributed in terms of age against salary. It is color-coded based on the age groups.

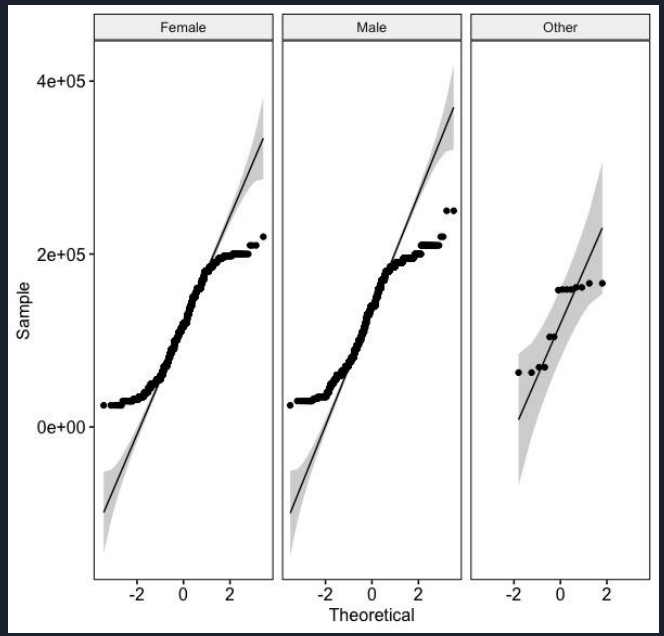

The box plot graph helps us visualize any outliers in each age group. It also helps us to decide whether the data is highly significant or not.

The scatter plot helps visualize the correlation between given age data and normal distribution on the plot, all the data falls along the reference line, confirming normal distribution

# Graphs (Gender)

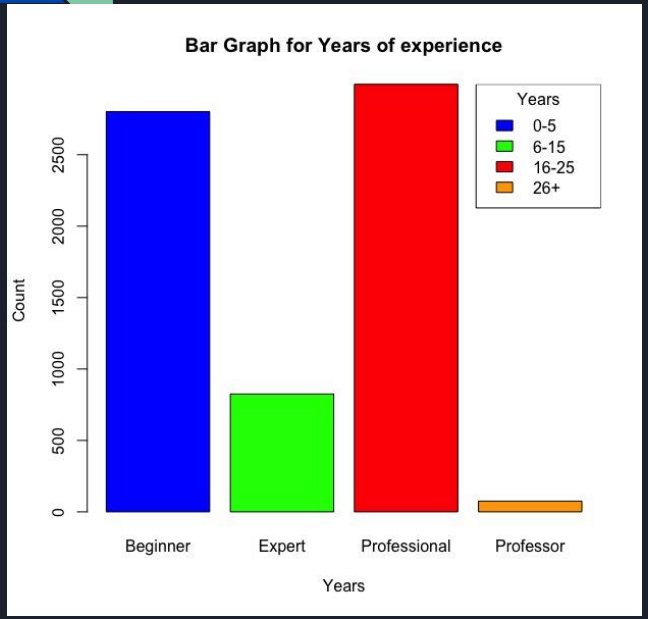




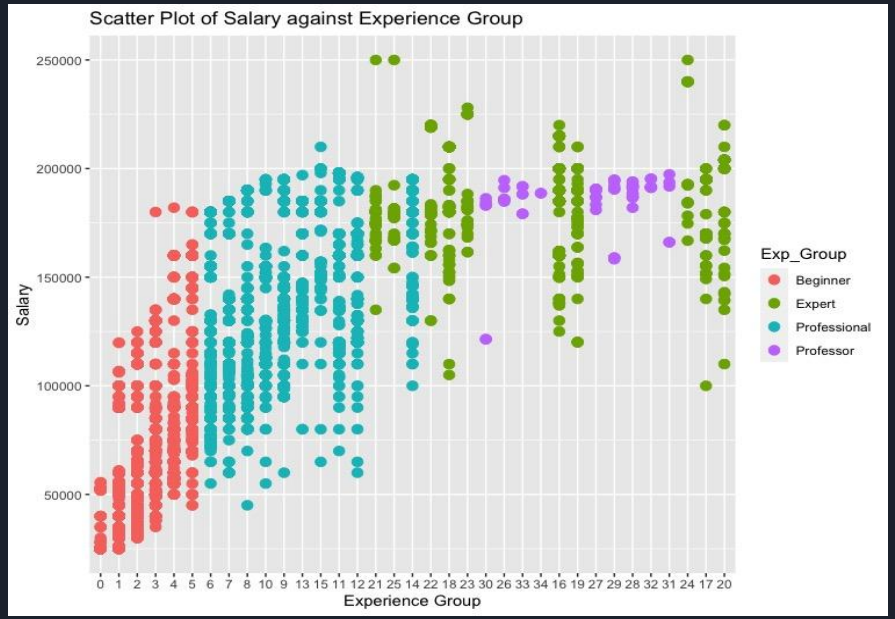



This graph shows the frequency of each Gender group. The box plot graph helps us visualize any outliers in each gender group which does not show any outliers.

# Graphs (Experience Level)

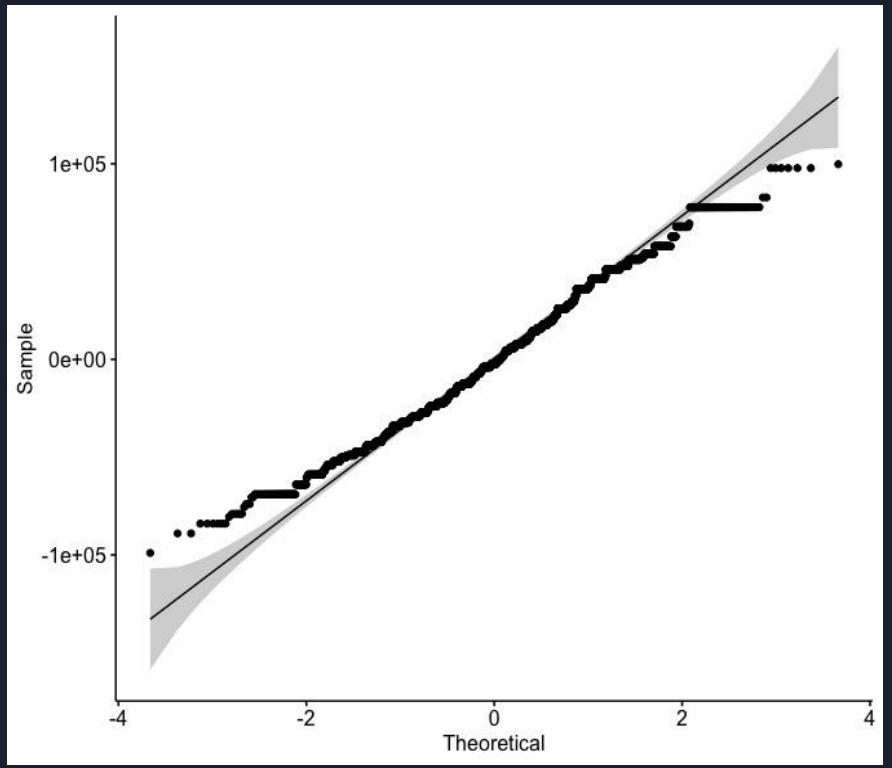



This graph shows the frequency of each experience levels within the data so we can better understand how the data is distributed in terms of years of experience.



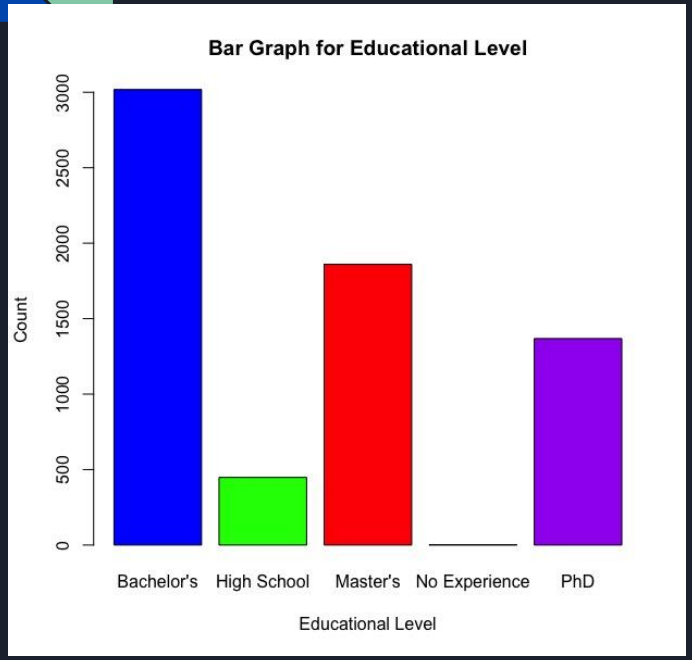

For this graph, we hope to see how the data is distributed in terms of experience levels against salary. It is color-coded based on the age groups.


The box plot graph helps us visualize any outliers in each experience level. It also helps us to decide whether the data is highly significant or not.

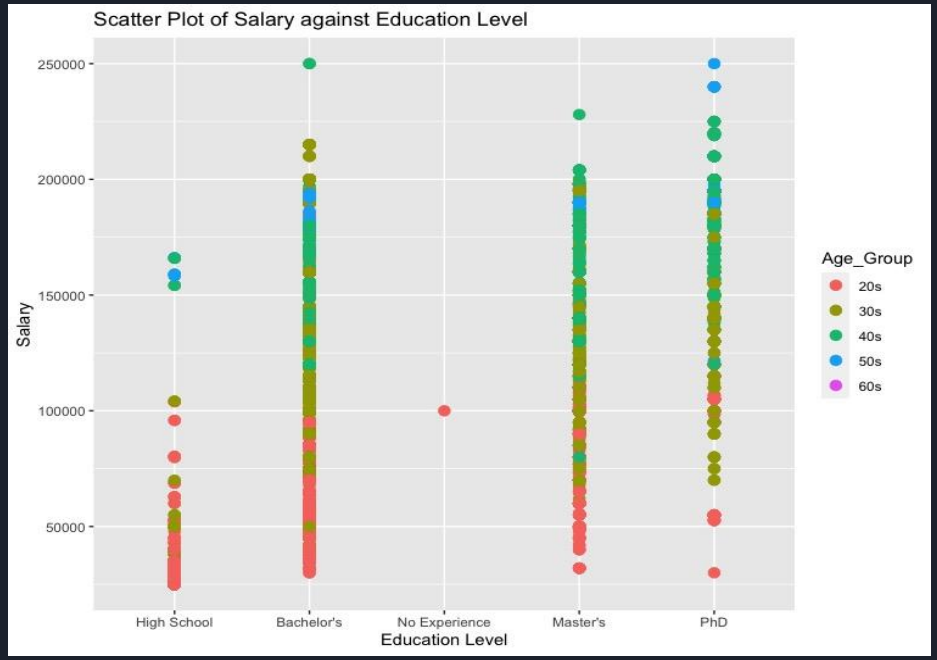



The scatter plot helps visualize the correlation between given years of experience data and normal distribution on the plot, all the data falls along the reference line, confirming normal distribution

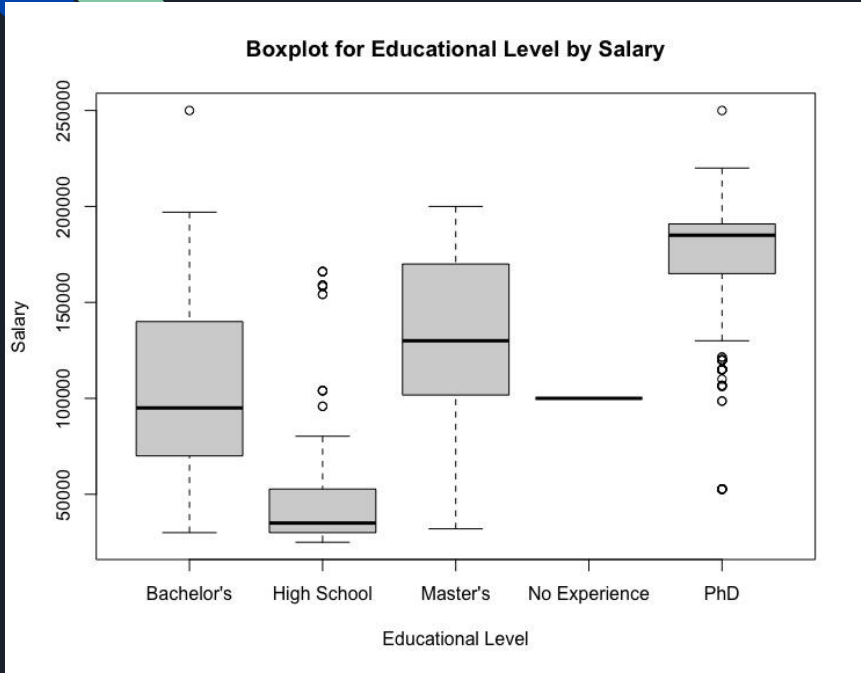
# Graphs (Education Level)



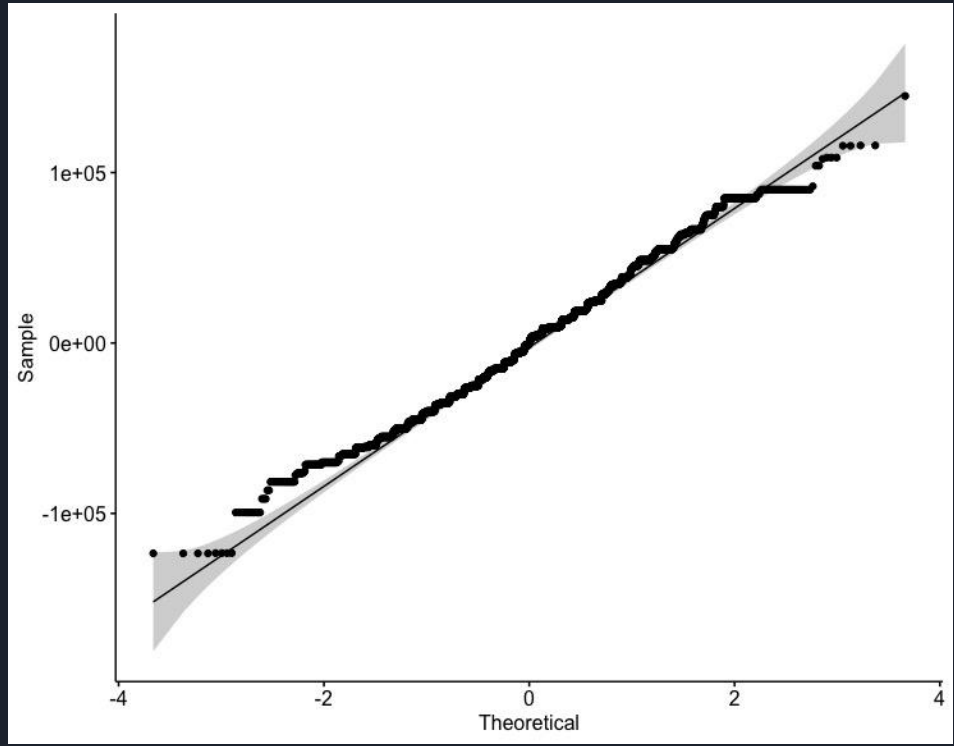

This graph shows the frequency of each Education level within the data so we can better understand how the data is distributed in terms of education level.




For this graph, we hope to see how the data is distributed in terms of education level against salary. It is color-coded based on the age groups.




The box plot graph helps us visualize any outliers in each Educational level. It also helps us to decide whether the data is highly significant or not.




The scatter plot helps visualize the correlation between given educational level data and normal distribution on the plot, all the data falls along the reference line, confirming normal distribution

# Interpretation

- For the variable where we compared two means, we used an independent t-test. However, for the other variables that have more than two means to compare such as average means across education levels we used the one way ANOVA tests.
- Positive correlation between the factors and salary (such as age and salary or gender and salary)
- The alternative hypothesis for each of the factors were correct as there was difference in the mean salaries between the populations within each variable.
- In the future, when financial decisions are made or decisions regarding salaries, we can take into account that various factors influence salary such as age, gender, experience, and education.

# End