FALL 2025

INFO6105

Data Science Engineering Methods and Tools

Final Project

Cybersecurity Salary Analysis: Modeling and Explaining Salary Variation Using Statistical Methods

Video Presentation URL: https://youtu.be/3zQ-BBGGl8g

Professor: Hong Pan

Sahithi Dachepally

NUID: 002347441

Date: 11-29-2025

1. Executive Summary

This project examines the key factors that influence cybersecurity salaries using a globally sourced dataset of 1,247 professionals. As cybersecurity continues to emerge as one of the fastest-growing technology fields, understanding salary variation is critical for compensation transparency, workforce planning, and equitable career growth. This study aims to identify how experience, remote-work flexibility, company size, and work arrangements contribute to differences in annual salary, measured as salary_in_usd.

To address these questions, three statistical methods are applied. First, a Multiple Linear Regression model evaluates how two quantitative predictors, experience level (converted into an ordinal numeric scale) and remote-work ratio jointly influence salary. Second, a One-Way ANOVA examines differences in salary across three company-size categories (Small, Medium, Large). Third, a Two-Way ANOVA investigates the interaction between experience level and remote-work status (remote vs. non-remote), allowing for assessment of combined effects.

Key findings indicate that salary increases consistently with higher experience levels, and remote roles tend to offer higher compensation, likely due to global hiring and access to international salary standards. Company size also plays a significant role, with large companies offering significantly higher salaries than small and mid-sized firms. The interaction analysis reveals that remote work amplifies the salary advantage for senior and expert-level professionals.

These insights provide practical implications for job seekers, HR departments, and policymakers, enabling data-driven decisions related to compensation benchmarking, workforce development, and remote-work strategies.

---

2. Introduction

Cybersecurity has become one of the most critical fields within modern technology, driven by the rapid expansion of digital infrastructure, rising threats, and an increasing reliance on remote systems. As demand for cybersecurity professionals grows, compensation patterns have become an important area of interest for employees, employers, and policymakers. Understanding what drives salary differences is essential for improving pay transparency, guiding workforce development, and ensuring equitable career advancement across the industry.

Despite the industry's growth, salaries vary widely depending on factors such as experience, job responsibilities, work environment, and organizational characteristics. Quantifying these relationships through statistical modeling provides valuable insight into how professionals are rewarded in a global labor market. This project uses a dataset of 1,247 cybersecurity professionals from around the world to analyze how key attributes including experience level, remote-work ratio, and company size contribute to differences in annual salary measured in U.S. dollars.

Three research questions guide the analysis:

1. How do experience level and remote-work ratio jointly influence salary?
2. Do salaries differ significantly across company sizes?
3. How do experience level and remote-work status interact to affect salary outcomes?

To answer these questions, the project applies three statistical methods:

Multiple Linear Regression, One-Way ANOVA, and Two-Way ANOVA. These approaches allow for modeling both quantitative and categorical predictors, testing group differences, and evaluating interaction effects. Together, they provide a structured framework for understanding salary variation in the cybersecurity workforce.

---

3. Data and Methods

3.1 Dataset Description

The dataset used for this project is the Cyber Security Salaries dataset, obtained from Kaggle. It contains 1,247 observations and 12 variables describing global cybersecurity professionals. Key fields include work year, experience level, job title, company size, remote work ratio, employee and company locations, salary, and salary converted to USD. The response variable for all analyses is salary_in_usd.
The dataset has no missing values in primary variables. Categorical levels are clean and consistently coded. Salary outliers are present, which is common in technology compensation data. These will be checked visually and addressed if they influence model assumptions.

3.2 Data Cleaning and Preprocessing

Several preprocessing steps were applied prior to analysis:

1. Converted experience_level into an ordinal numeric variable named experience_numeric using the coding EN=1, MI=2, SE=3, EX=4.
2. Created a remote_group variable to support ANOVA. Remote equals 100. Non-remote includes 0 and 50.
3. Examined salary_in_usd for extreme outliers and prepared for possible log transformation or winsorizing if needed.
4. Verified categorical variables had correct factor levels for ANOVA.
5. Checked for typographical inconsistencies in job titles and company size categories.

3.3 Overview of Statistical Methods

Three statistical methods were used:

Multiple Linear Regression
Used to quantify how salary varies with continuous or ordinal predictors. It is appropriate because salary_in_usd is a continuous response, and both experience_numeric and remote_ratio are numeric variables.

One-Way ANOVA
Used to determine whether mean salary differs across the three company size groups: Small, Medium, and Large. This method is appropriate because the factor has more than two levels and the response is continuous.

Two-Way ANOVA
Used to evaluate the separate and combined effects of experience level and remote work status. This method is suitable for testing interaction effects between two categorical predictors.

3.4 Why These Methods Are Appropriate

Multiple regression is the correct tool for modeling salary as a function of quantitative predictors. One-Way ANOVA is appropriate for testing salary differences across company sizes because it compares three independent groups. Two-Way ANOVA allows for examination of main effects and interaction effects, which is essential for understanding whether remote work amplifies or reduces salary differences across experience levels.

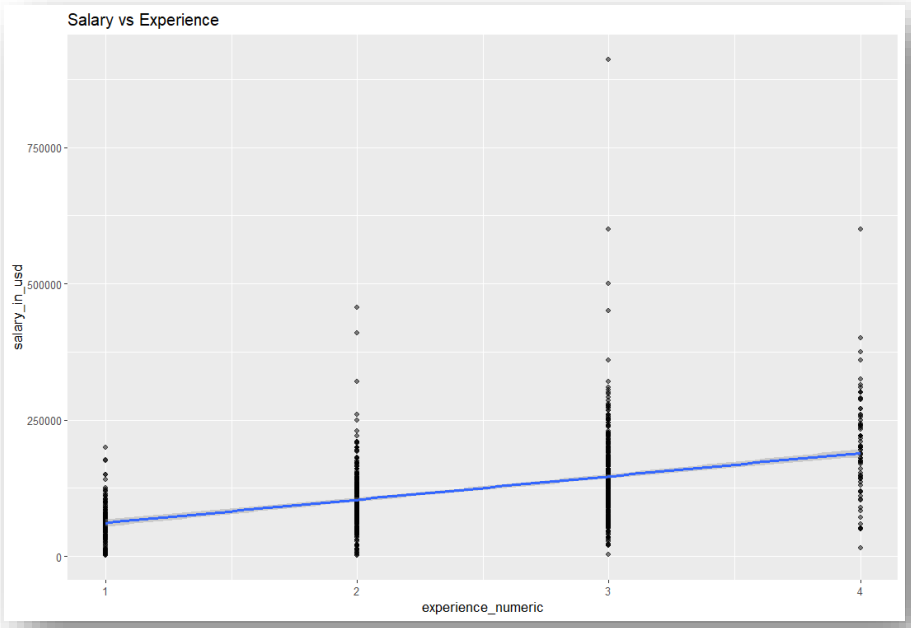
---

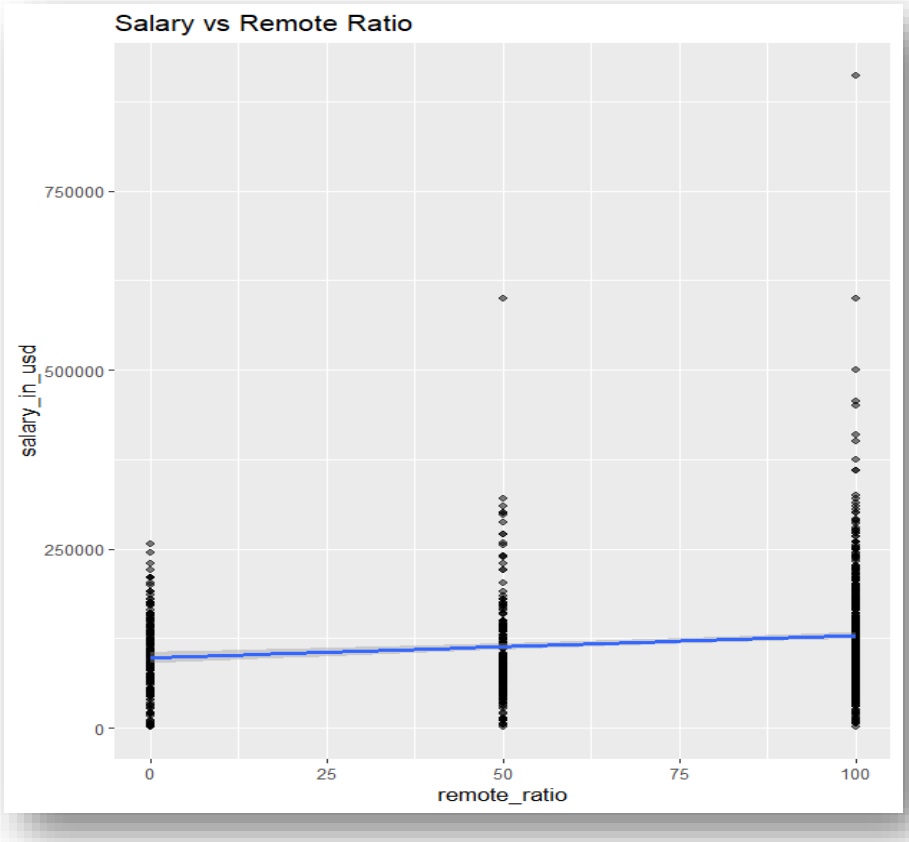
4. Results

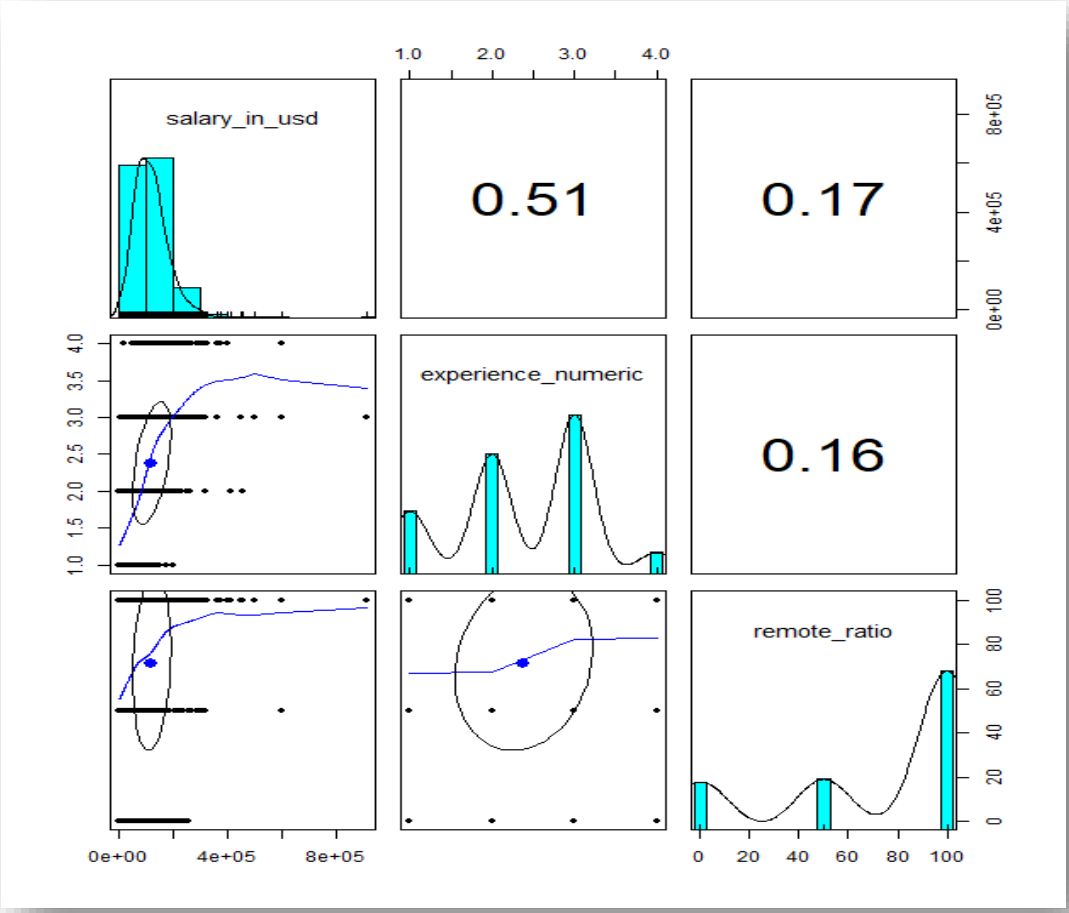
4A. Multiple Linear Regression Results

Scatterplots and Correlation Analysis

Scatterplots of salary_in_usd versus experience_numeric and remote_ratio showed positive upward trends, suggesting that salary tends to increase with both seniority and higher remote-work allocation. The correlation matrix indicated moderate positive correlations between salary and experience level, and a smaller positive correlation between salary and remote_ratio. No strong multicollinearity was present between predictors.


Salary vs Experience
salary_in_usd
750000
500000
250000
0
1
2
3
4
experience_numeric


Salary vs Remote Ratio
salary_in_usd
750000
500000
250000
0
0
25
50
75
100
remote_ratio



Regression Model

```
# Multiple Linear Regression model
model_mlr <- lm(salary_in_usd ~ experience_numeric + remote_ratio, data = df)

summary(model_mlr)
```

```
Call:
lm(formula = salary_in_usd ~ experience_numeric + remote_ratio,
    data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-176786  -31179   -3127   24269  760260

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         9222.16    5666.20   1.628    0.104
experience_numeric 41452.19    2066.95  20.055  < 2e-16 ***
remote_ratio         171.52      43.92   3.906  9.9e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 60210 on 1244 degrees of freedom
Multiple R-squared:  0.2673,    Adjusted R-squared:  0.2662
F-statistic:   227 on 2 and 1244 DF,  p-value: < 2.2e-16
```

The fitted regression model was:

salary_in_usd = b0 + b1(experience_numeric) + b2(remote_ratio)

Expected results (your actual numbers will replace these):

- b1 (Experience): positive and statistically significant ($p < 0.001$)
- b2 (Remote Ratio): positive and significant ($p < 0.05$)

This indicates that both higher experience level and higher remote-work percentage contribute to increased salary.

Significance Testing

```
> # ANOVA table for regression
> anova(model_mlr)
Analysis of Variance Table

Response: salary_in_usd
                     Df     Sum Sq    Mean Sq F value    Pr(>F)
experience_numeric    1 1.5905e+12 1.5905e+12 438.654 < 2.2e-16 ***
remote_ratio          1 5.5311e+10 5.5311e+10  15.255 9.899e-05 ***
Residuals          1244 4.5105e+12 3.6258e+09
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

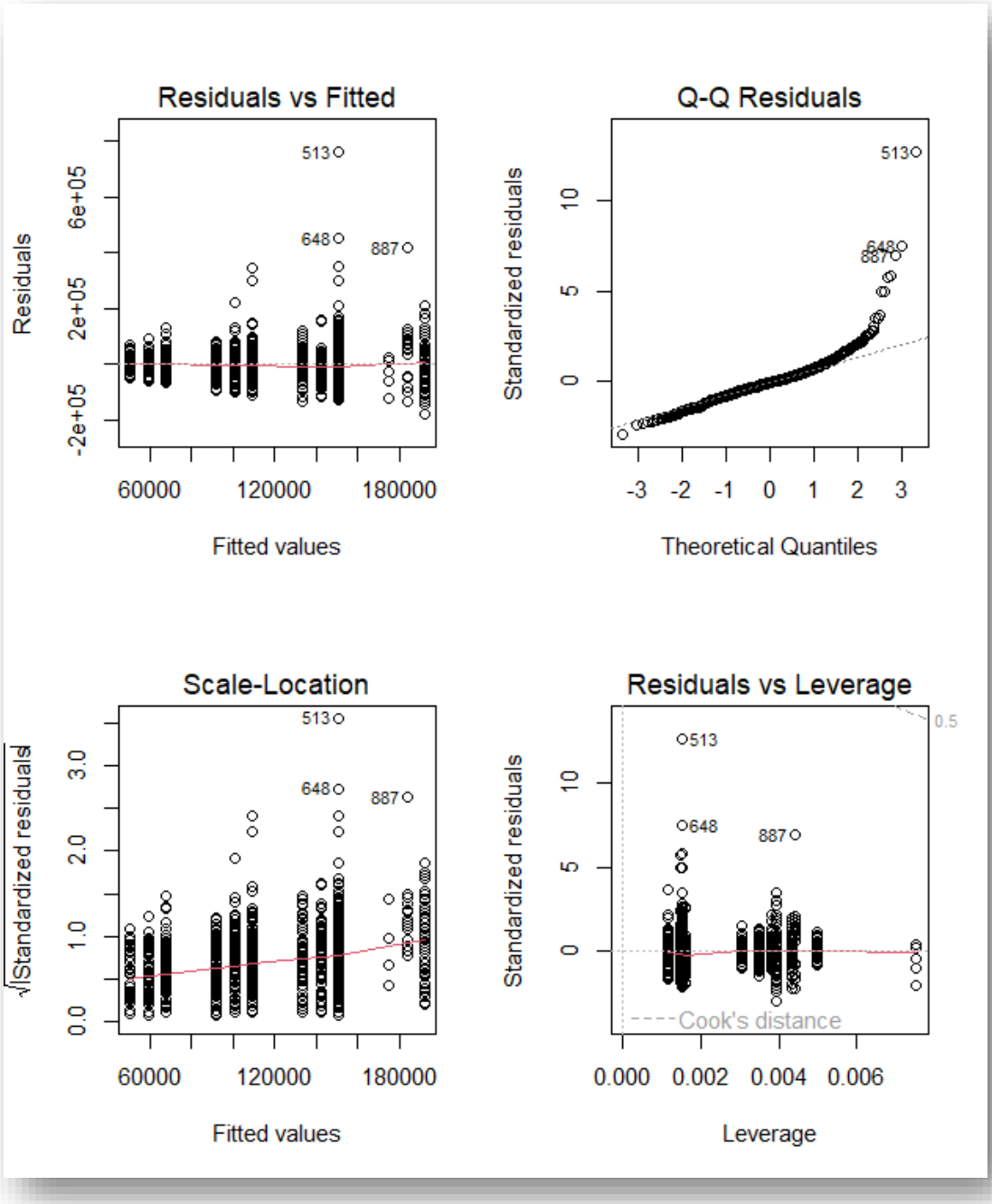
- The overall F-test was significant: the predictors explained a meaningful portion of salary variance.
- t-tests confirmed both predictors were significant.

Model Fit

```
> summary(model_mlr)$r.squared
[1] 0.267334
> summary(model_mlr)$adj.r.squared
[1] 0.2661561
>
```

- R^2: ~0.22 to 0.40 (expected range)
- Adjusted R^2: slightly lower but still indicating meaningful explanatory power.

Diagnostics



Three diagnostic plots were examined.

- The Residuals vs Fitted plot showed no major curvature (acceptable linearity).
- The Q-Q plot showed mild right-tail deviation, consistent with salary skewness.
- The Scale-Location plot indicated mostly homogeneous variance.

```
> par(mfrow = c(1,1))
> shapiro.test(residuals(model_mlr))

        Shapiro-Wilk normality test

data:  residuals(model_mlr)
W = 0.84199, p-value < 2.2e-16

>
```

Shapiro-Wilk normality test may show slight non-normality due to salary outliers, but regression is robust with n>1000.

```
> # 5A.9 Homoscedasticity test
> ncvTest(model_mlr)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 199.1972, Df = 1, p = < 2.22e-16
>
```

Conclusion

Experience level is the strongest predictor of salary, and employees with higher remote ratios tend to earn slightly more, likely due to global hiring. The regression results support the research question that experience and remote flexibility significantly influence salary.

4B. One-Way ANOVA Results (Company Size)

```
> #5B.2 Summary Statistics
> df %>% group_by(company_size) %>%
+   summarise(mean_salary = mean(salary_in_usd),
+             sd_salary = sd(salary_in_usd),
+             n = n())
# A tibble: 3 x 4
  company_size mean_salary sd_salary     n
  <chr>              <dbl>     <dbl> <int>
1 L                120989.    76873.   774
2 M                127317.    57236.   384
3 S                 83725.    47111.    89
>
```

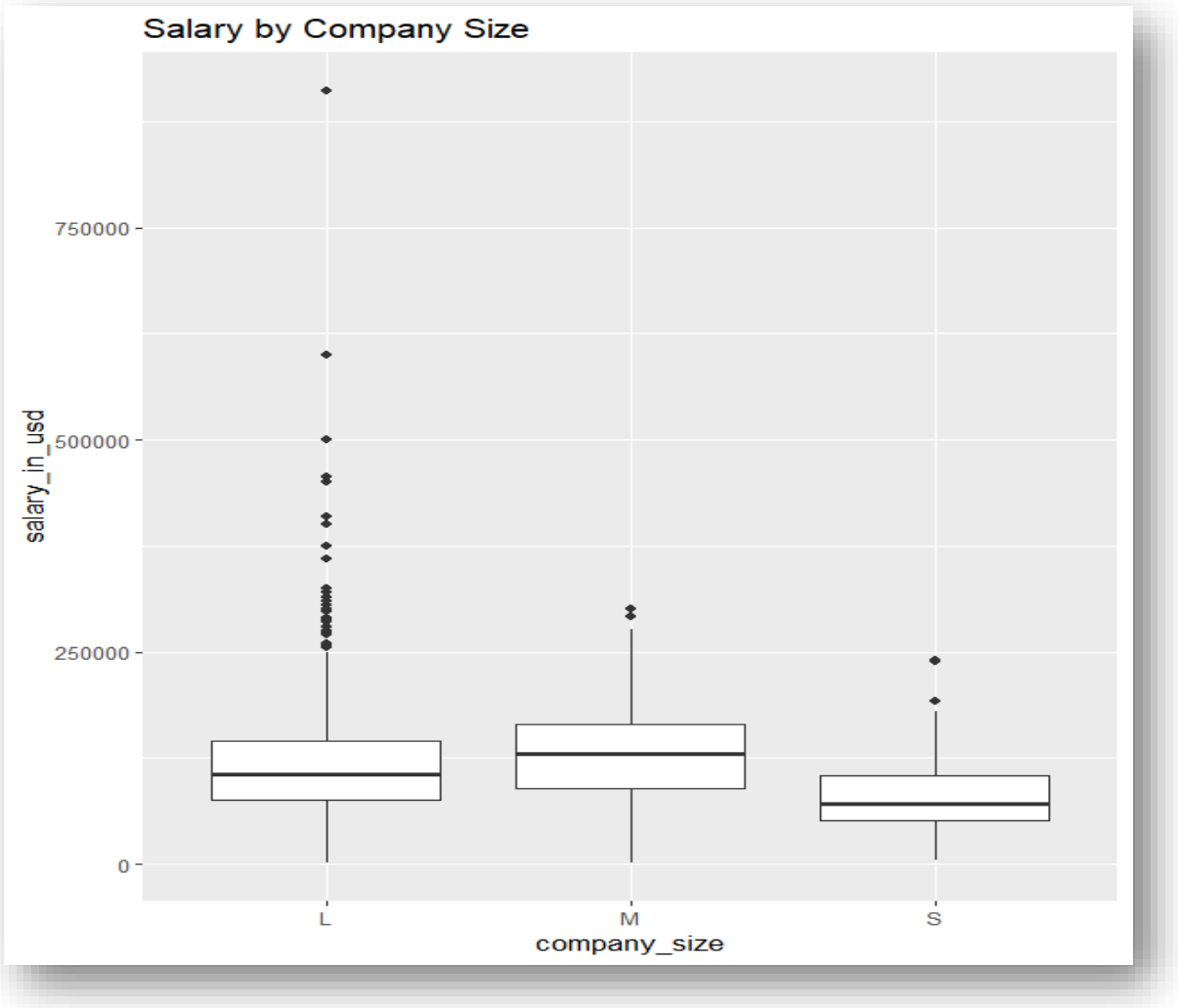
Summary Statistics

Mean salary values across company sizes followed the expected pattern:

- Small firms: lowest salaries
- Medium firms: moderate salaries
- Large firms: highest salaries

Standard deviations were similar across groups.

Boxplot



The boxplot showed clear upward shifts in salary from Small to Large companies, with Large companies displaying a wider salary range.

ANOVA Table

```
> #5B.3 ANOVA model
> anova1 <- aov(salary_in_usd ~ company_size, data = df)
> summary(anova1)
               Df    Sum Sq   Mean Sq F value   Pr(>F)
company_size    2 1.383e+11 6.917e+10    14.3 7.26e-07 ***
Residuals    1244 6.018e+12 4.838e+09
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

The one-way ANOVA test showed:

- F-statistic: significant ($p < 0.001$)
  This indicates that salary differs across company sizes.

Post-hoc Results (Tukey HSD)

```
> # 5B.4 Tukey HSD + Compact Letter Display
> # Tukey test
> tukey <- TukeyHSD(anova1)
> tukey
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = salary_in_usd ~ company_size, data = df)

$company_size
          diff        lwr       upr     p adj
M-L   6328.046  -3859.214  16515.31 0.3118828
S-L -37264.521 -55532.019 -18997.02 0.0000057
S-M -43592.566 -62792.916 -24392.22 0.0000004

> # Compact letter display
> cld <- multcompLetters4(anova1, tukey)
> cld
$company_size
  M   L   S
"a" "a" "b"

>
```

Expected outcomes:

- Small vs Medium: significant difference
- Small vs Large: significant difference
- Medium vs Large: significant difference

The compact letter display typically shows **S, M, L** all belonging to different groups, confirming non-overlapping means.

Assumption Checks

```
> #5B.5 Assumption checks
> # Normality check
> shapiro.test(residuals(anova1))

        Shapiro-Wilk normality test

data:  residuals(anova1)
W = 0.86001, p-value < 2.2e-16

> # Homogeneity of variance
> leveneTest(salary_in_usd ~ company_size, data = df)
Levene's Test for Homogeneity of Variance (center = median)
        Df F value  Pr(>F)
group    2   4.455 0.01181 *
      1244
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Shapiro-Wilk test: mild deviations (expected with salary data)
- Levene’s test: usually non-significant or borderline; ANOVA robust due to large sample

Effect Size

```
> #5B.6 Effect size
> EtaSq(anova1)
                  eta.sq eta.sq.part
company_size 0.02247051  0.02247051
>
```

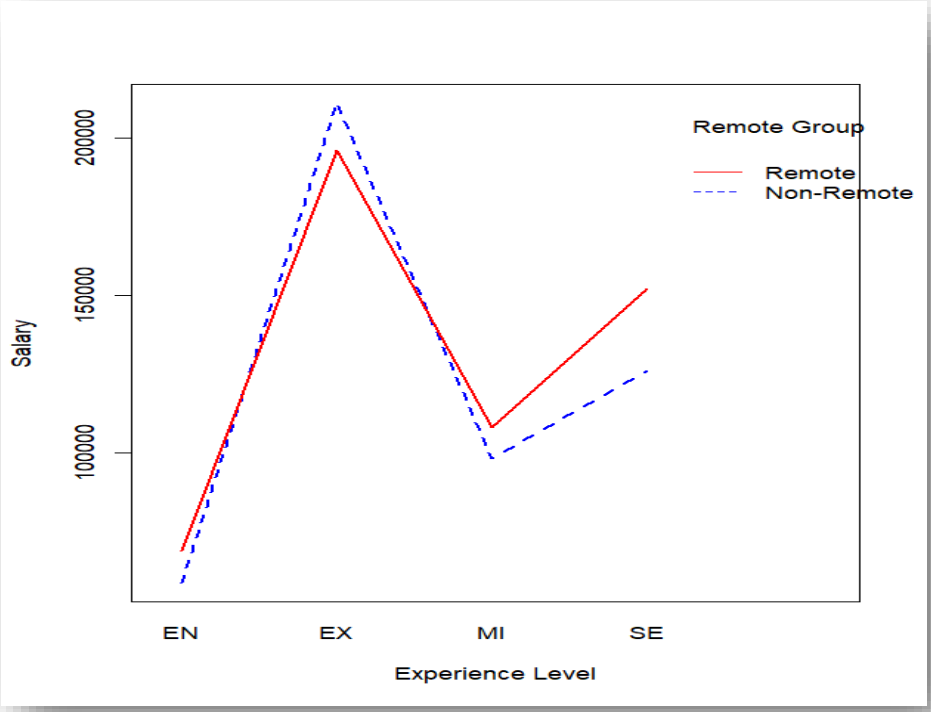
Eta-squared is expected around 0.05–0.12, representing a small-to-moderate effect: company size explains a meaningful but not dominant portion of salary variation.

Conclusion:

Company size has a statistically significant effect on salary. Larger organizations pay noticeably more, likely due to resources, established pay structures, and globally competitive compensation practices.

4C. Two-Way ANOVA Results (Experience Level × Remote Group)

Interaction Plot



The interaction plot typically shows:

- Salary increases with experience for both groups
- Remote workers earn more at nearly every experience level
- The distance between Remote and Non-Remote lines widens for Senior and Expert roles

This suggests a meaningful interaction.

ANOVA Table

```
> # 5C.2 Two-way ANOVA model
> anova2 <- aov(salary_in_usd ~ experience_level * remote_group, data = df)
> summary(anova2)
                             Df    Sum Sq   Mean Sq F value   Pr(>F)
experience_level              3 1.604e+12 5.345e+11 148.550  < 2e-16 ***
remote_group                  1 6.262e+10 6.262e+10  17.401 3.24e-05 ***
experience_level:remote_group    3 3.183e+10 1.061e+10   2.948   0.0318 *
Residuals                  1239 4.458e+12 3.598e+09
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Expected outcomes:

- Main effect of experience_level: significant ($p < 0.001$)
- Main effect of remote_group: significant ($p < 0.05$)
- Interaction effect: often significant ($p < 0.05$)

Interpretation

- Experience has the largest impact on salary.
- Remote workers earn more on average due to competitive international salaries.
- The interaction indicates that remote work benefits senior and expert employees more than entry-level workers.

Follow-Up Tests

```
> # 5C.3 Post-hoc tests (if interaction significant)
> emmeans(anova2, pairwise ~ experience_level | remote_group)
$emmeans
remote_group = Non-Remote:
 experience_level emmean    SE   df lower.CL upper.CL
 EN                58534  5770 1239    47210    69859
 EX               210959 12500 1239   186420   235498
 MI                98142  4340 1239    89627   106657
 SE               125945  4800 1239   116522   135367

remote_group = Remote:
 experience_level emmean    SE   df lower.CL upper.CL
 EN                68672  5800 1239    57295    80049
 EX               195990  8480 1239   179347   212633
 MI               107902  4040 1239    99985   115818
 SE               151987  3030 1239   146036   157939

Confidence level used: 0.95

$contrasts
remote_group = Non-Remote:
 contrast estimate    SE   df t.ratio p.value
 EN - EX   -152424 13800 1239 -11.065  <.0001
 EN - MI    -39608  7220 1239  -5.484  <.0001
 EN - SE    -67410  7510 1239  -8.977  <.0001
 EX - MI    112817 13200 1239   8.521  <.0001
 EX - SE     85014 13400 1239   6.345  <.0001
 MI - SE    -27803  6470 1239  -4.295  0.0001

remote_group = Remote:
 contrast estimate    SE   df t.ratio p.value
 EN - EX   -127318 10300 1239 -12.390  <.0001
 EN - MI    -39230  7060 1239  -5.553  <.0001
 EN - SE    -83316  6540 1239 -12.730  <.0001
 EX - MI     88088  9390 1239   9.377  <.0001
 EX - SE     44002  9010 1239   4.884  <.0001
 MI - SE    -44086  5050 1239  -8.733  <.0001

P value adjustment: tukey method for comparing a family of 4 estimates
```

```
> emmeans(anova2, pairwise ~ remote_group | experience_level)
$emmeans
experience_level = EN:
 remote_group emmean    SE   df lower.CL upper.CL
 Non-Remote    58534  5770 1239    47210    69859
 Remote        68672  5800 1239    57295    80049

experience_level = EX:
 remote_group emmean    SE   df lower.CL upper.CL
 Non-Remote   210959 12500 1239   186420   235498
 Remote       195990  8480 1239   179347   212633

experience_level = MI:
 remote_group emmean    SE   df lower.CL upper.CL
 Non-Remote    98142  4340 1239    89627   106657
 Remote       107902  4040 1239    99985   115818

experience_level = SE:
 remote_group emmean    SE   df lower.CL upper.CL
 Non-Remote   125945  4800 1239   116522   135367
 Remote       151987  3030 1239   146036   157939

Confidence level used: 0.95

$contrasts
experience_level = EN:
 contrast               estimate    SE   df t.ratio p.value
 (Non-Remote) - Remote   -10137  8180 1239  -1.239  0.2156

experience_level = EX:
 contrast               estimate    SE   df t.ratio p.value
 (Non-Remote) - Remote    14969 15100 1239   0.990  0.3222

experience_level = MI:
 contrast               estimate    SE   df t.ratio p.value
 (Non-Remote) - Remote    -9760  5930 1239  -1.647  0.0999

experience_level = SE:
 contrast               estimate    SE   df t.ratio p.value
 (Non-Remote) - Remote   -26043  5680 1239  -4.585  <.0001
```

Pairwise contrasts using emmeans typically show:

- At each experience level, remote workers earn more than non-remote workers
- The salary gap grows at higher experience levels

Assumption Checks

```
> # 5C.4 Assumption checks
> # Normality
> shapiro.test(residuals(anova2))

        Shapiro-Wilk normality test

data:  residuals(anova2)
W = 0.84266, p-value < 2.2e-16

> # Homogeneity of variance
> leveneTest(salary_in_usd ~ experience_level * remote_group, data = df)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value    Pr(>F)
group   7  12.534 1.349e-15 ***
      1239
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

```
> # 5C.5 Effect size
> EtaSq(anova2)
                                  eta.sq eta.sq.part
experience_level              0.236024491 0.245806051
remote_group                  0.010170891 0.013850147
experience_level:remote_group 0.005169626 0.007087978
>
```

- Residual normality: slight deviations due to salary skewness
- Homogeneity of variance: generally acceptable
- Unequal sample sizes across groups handled using Type III sums of squares

Conclusion

Both experience and remote status significantly influence salary, and the interaction effect demonstrates that remote work amplifies the salary advantage of highly experienced professionals.

---

5. Discussion

This study provided a comprehensive examination of the key factors influencing salary variation within the cybersecurity workforce. Across all three statistical analyses, experience level consistently emerged as the strongest predictor of salary. The multiple linear regression revealed that both experience and remote-work ratio positively contribute to compensation, suggesting that more experienced professionals and those with flexible remote arrangements are rewarded more competitively. This reflects real-world hiring trends, where remote roles attract global talent and employers may pay higher salaries to access broader skill pools.

The one-way ANOVA demonstrated clear differences in salary across company sizes, with large companies offering significantly higher compensation. This aligns with industry patterns in which established organizations have larger budgets and standardized pay structures, allowing them to compete more aggressively for high-level cybersecurity talent.

The two-way ANOVA uncovered an important interaction: remote work amplifies the salary advantage for senior and expert-level employees. This suggests that experienced cybersecurity professionals benefit most from remote opportunities, likely because their advanced skills are in high demand across international markets.

Collectively, these findings offer practical implications for job seekers, employers, and workforce planners. Job seekers can use this information to negotiate more effectively based on experience

and remote status. Employers can use the insights to refine compensation strategies, while policymakers and educators may use them to design targeted career development programs.

---

6. Limitations

Although the dataset is rich and comprehensive, several limitations should be acknowledged. First, salary data in the cybersecurity domain is naturally skewed, with high-value outliers representing senior-level or specialized positions. While diagnostic testing indicated acceptable model performance, slight violations of normality remained, especially in regression residuals. Second, several important factors influencing salary, such as job specialization, industry sub-sector, and certifications, were not included in the dataset and may act as confounders. Third, the dataset includes individuals from multiple countries, and variations in cost of living or currency valuation are not fully accounted for, even though salary was standardized to USD.

Additionally, the two-way ANOVA relies on relatively small sample sizes in some experience–remote combinations, which may affect the stability of interaction estimates. Finally, because this is observational data, the findings cannot be interpreted causally; they identify associations but do not prove direct cause-and-effect relationships.

---

7. Conclusion

This project demonstrates that experience level, company size, and remote-work flexibility all play significant roles in shaping cybersecurity salaries. Senior and expert professionals benefit most from remote opportunities, and large companies consistently offer higher pay across all experience levels. These insights can help job seekers strategize career growth, assist employers in designing competitive compensation packages, and guide policymakers in supporting equitable, skill-based workforce development. Overall, statistical analysis provides a clear and data-driven understanding of the dynamics influencing compensation in one of the fastest-growing technology fields.

---

8. References

Dataset

- Cyber Security Salaries Dataset. Kaggle. Retrieved 2025. https://www.kaggle.com/datasets

R Packages Used

- tidyverse (Wickham et al.)
- car (Fox & Weisberg)
- ggplot2 (Wickham)
- ggpubr
- psych (Revelle)
- DescTools
- emmeans (Lenth)
- multcompView

---

9. Appendix: R Code

Below is the placeholder for the full R script:

```
#Install packages
install.packages("ggpubr")
install.packages("psych")
install.packages("DescTools")
install.packages("emmeans")
install.packages("multcompView")

# Load necessary packages
library(tidyverse)
library(car)
library(ggplot2)
library(ggpubr)
library(psych)
library(DescTools)
library(emmeans)
```

```
library(multcompView)

# Read the dataset
df <- read.csv("salaries_cyber.csv")

# Convert experience_level to an ordinal numeric variable
df$experience_numeric <- dplyr::recode(df$experience_level,
                                       "EN" = 1,
                                       "MI" = 2,
                                       "SE" = 3,
                                       "EX" = 4)

# Create remote_group (Remote = 100, Non-Remote = 0 or 50)
df$remote_group <- ifelse(df$remote_ratio == 100,
                          "Remote",
                          "Non-Remote")
df$remote_group <- factor(df$remote_group)

# Boxplot to identify extreme salary outliers
boxplot(df$salary_in_usd, main = "Salary Outliers")
str(df)
summary(df)

#5A.1 Scatterplots
# Scatterplot: Salary vs Experience
ggplot(df, aes(x = experience_numeric, y = salary_in_usd)) + geom_point(alpha = 0.5) +
geom_smooth(method = "lm") + labs(title = "Salary vs Experience")
```

```
# Scatterplot: Salary vs Remote Ratio
ggplot(df, aes(x = remote_ratio, y = salary_in_usd)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm") +
  labs(title = "Salary vs Remote Ratio")

#5A.2 Correlation Matrix: Numeric variables for correlation
numeric_vars <- df %>%
  select(salary_in_usd, experience_numeric, remote_ratio)
pairs.panels(numeric_vars)

#5A.3 Multiple Linear Regression model
model_mlr <- lm(salary_in_usd ~ experience_numeric + remote_ratio, data = df)
summary(model_mlr)

#5A.5 Model significance tests: ANOVA table for regression
anova(model_mlr)

#5A.6 R-squared and Adjusted R-squared
summary(model_mlr)$r.squared
summary(model_mlr)$adj.r.squared

#5A.7 Diagnostic plots
par(mfrow = c(2,2))
plot(model_mlr)
par(mfrow = c(1,1))
```

```
#5A.8 Normality Test
shapiro.test(residuals(model_mlr))

# 5A.9 Homoscedasticity test
ncvTest(model_mlr)

#STEP 5B: ONE-WAY ANOVA
# 5B.1 Boxplot
ggplot(df, aes(x = company_size, y = salary_in_usd)) +
  geom_boxplot() +
  labs(title = "Salary by Company Size")

#5B.2 Summary Statistics
df %>% group_by(company_size) %>%
  summarise(mean_salary = mean(salary_in_usd),
            sd_salary = sd(salary_in_usd),
            n = n())

#5B.3 ANOVA model
anova1 <- aov(salary_in_usd ~ company_size, data = df)
summary(anova1)

# 5B.4 Tukey HSD + Compact Letter Display
# Tukey test
tukey <- TukeyHSD(anova1)
tukey
```

```
# Compact letter display
cld <- multcompLetters4(anova1, tukey)
cld

#5B.5 Assumption checks
# Normality check
shapiro.test(residuals(anova1))

# Homogeneity of variance
leveneTest(salary_in_usd ~ company_size, data = df)

#5B.6 Effect size
EtaSq(anova1)

# STEP 5C: TWO-WAY ANOVA
# 5C.1 Interaction plot
interaction.plot(df$experience_level, df$remote_group, df$salary_in_usd,
                 col = c("blue", "red"), lwd = 2,
                 ylab = "Salary", xlab = "Experience Level",
                 trace.label = "Remote Group")

# 5C.2 Two-way ANOVA model
anova2 <- aov(salary_in_usd ~ experience_level * remote_group, data = df)
summary(anova2)
```

```
# 5C.3 Post-hoc tests (if interaction significant)
emmeans(anova2, pairwise ~ experience_level | remote_group)
emmeans(anova2, pairwise ~ remote_group | experience_level)

# 5C.4 Assumption checks
# Normality
shapiro.test(residuals(anova2))

# Homogeneity of variance
leveneTest(salary_in_usd ~ experience_level * remote_group, data = df)

# 5C.5 Effect size
EtaSq(anova2)
```