



פרויקט נושאים נבחרים בסטטיסטיקה

המחלקה להנדסת תעשיה וניהול אוניברסיטת בן גוריון



תוכן עניינים

[תקציר 3](#_Toc127824230)

[שאלות המחקר/ השערות המחקר 3](#_Toc127824231)

[מבוא 3](#_Toc127824232)

[שיטה 4](#_Toc127824233)

[נתונים 4](#_Toc127824234)

[תיאור והסבר מפורט של המשתנים במחקר הסטטיסטי 5](#_Toc127824235)

[המודל בו השתמשתם 9](#_Toc127824236)

[14](#_Toc127824237)

[דיון ומסקנות 15](#_Toc127824238)

[נספחים 16](#_Toc127824239)



# 

# תקציר

המחקר הנוכחי אוסף נתונים על שכר עבודה בתחום מדעי הנתונים, בניה מלאכותית, מערכות לומדות וכדו, ובוחנים את הגורמים המשפיעים על הפיצויים בתחום זה, כגון סוג משרה, רמת ניסון בתפקיד, היקף משרה, גול משרה שנעשה מרחוק, גודל חברה לפי מספר עובדים ואם העובד עובד באותה מדינה שגר בה. נבחן כל גורם לבד ואת הפיזור שלו. ויצא שהנתונים שלנו שכמעט כולם בעלי משרות מלאות לכן התעלמנו ממשתנה הזה במודל הרגרסיה, אנו מוצאים גם שרמת הניסיון הם המניעים העיקריים להפרשי השכר, כאשר אנשי מקצוע בעלי ניסיון רב יותר של שנים מחזיקים במשכורות גבוהות יותר.

לתוצאות המחקר הזה יש השלכות חשובות על חברות המבקשות למשוך ולשמר כישרונות מובילים בבינה מלאכותית ומדעי נתונים. על ידי הבנת הגורמים המשפיעים על השכר בתחום זה, ארגונים יכולים לעצב חבילות תגמול תחרותיות ולמצב את עצמם כמעסיקים אטרקטיביים לאנשי מקצוע מובילים בתחום הנתונים.

## שאלות המחקר/ השערות המחקר

* השפעת הגורמים השונים על השכר
* ההבדל בשכר בין משרות שונות של דאטה ואיך הם מושפעים משינויים במשתנים

מבוא

אתר <https://ai-jobs.net/> אוסף מידע על שכר באופן אנונימי מאנשי מקצוע בכל רחבי העולם בתחום AI/ML/Data Science והופך אותו לזמין לציבור עבור כל אחד להשתמש, לשתף ולשחק איתו.

המטרה העיקרית היא לקבל נתונים שיכולים לספק הדרכה טובה יותר לגבי מה שמקבלים תשלום ברחבי העולם. אז מתחילים, מקצוענים מנוסים, מנהלי גיוס, מגייסים וגם מייסדי סטארט-אפים או אנשים שרוצים לעבור קריירה יכולים לקבל החלטות מושכלות יותר.

הביקוש לאנשי מקצוע בבינה מלאכותית (AI) ומדעי נתונים נמצא בעלייה. בשנים האחרונות, ריבוי הנתונים הוביל לצורך הולך וגובר באנשי מקצוע מיומנים בתחום הנתונים שיכולים להפוך נתונים גולמיים לתובנות ניתנות לפעולה. כתוצאה מכך, חברות ברחבי תעשיות מבקשות לגייס כישרונות מובילים בתחומים של בינה מלאכותית ומדעי נתונים כדי לעזור להן לקבל החלטות מונעות נתונים, משיפור חוויות הלקוחות ועד לאופטימיזציה של תפעול והנעת חדשנות.

עם זאת, ככל שהביקוש לכישרון בינה מלאכותית ומדעי הנתונים גדל, כך גדלה התחרות על אנשי המקצוע הללו, מה שהופך את זה לחיוני לחברות להציע משכורות ושכר תחרותיים.

לאור הביקוש הגבוה והשכר התחרותי למשרות בינה מלאכותית ומדעי הנתונים, יש צורך להבין את הגורמים המשפיעים על התגמול בתחום זה. במאמר זה, אנו שואפים לחקור את המשכורות והשכר של אנשי מקצוע בתחום הבינה המלאכותית ומדעי הנתונים ולחקור את הגורמים המשפיעים על התגמול בתחום זה.

המחקר ישתמש בנתונים מאתר ai-jobs, כדי לנתח מגמות שכר עבור אנשי מקצוע בתחום הנתונים, תוך התמקדות בהבדלים בתגמול עבור משרות בינה מלאכותית ומדעי נתונים. כמו כן, נבחן את הגורמים המשפיעים על השכר והשכר, כגון מיקום העבודה, רמת הניסיון, ההשכלה והתעשייה.

בסך הכל, הממצאים של מחקר זה יספקו תובנות חשובות לחברות המבקשות למשוך ולשמר כישרונות מובילים בינה מלאכותית ומדעי נתונים. על ידי הבנת הגורמים המשפיעים על השכר בתחום זה, ארגונים יכולים לעצב חבילות תגמול תחרותיות המושכות ומשמרות את אנשי המקצוע הטובים ביותר בתחום הנתונים, ולמצב את עצמם כמובילים בתחום ה-AI ומדעי הנתונים.

שיטה :

נתונים – מאיפה נלקחו, איך בוצע הניסוי, מיהם הנבדקים אם ישנם וכו'..

|  |  |  |
| --- | --- | --- |
| שם משתנה | הסבר | סוג |
| salary | שכר שנתי באלפי דולרים | רציף |
| job\_title | תפקיד:   * Data Scientist * Data Engineer * Data Analyst * Data Architect * Machine Learning Engineer * Research Scientist | קטגוריאלי |
| experience\_level | רמת הניסיון בתפקיד:   * EN: Entry level/Junior * MI: Mid-level/Intermediate * SE: Senior-level/Expert * EX: Executive-level/Director | קטגוריאלי |
| employment\_type | סוג המשרה:   * PT: משרה חלקית * FT: משרה מלאה * CT: חוזה * FL: פרילנסר | קטגוריאלי |
| work\_year | שנת עבודה: 2020, 2021, 2022, 2023 | קטגוריאלי |
| remote\_ratio | כמות העבודה שנעשית מרחוק:   * 0: אין עבודה מרחוק * 50: עבודה מרחוק חלקית * 100: עבודה מרחוק | קטגוריאלי |
| company\_size | גודל חברה לפי מספר עובדים ממוצע במהלך השנה:   * S: פחות מ-50 * M: מ-50 ל-250 * L: יותר מ-250 | קטגוריאלי |
| same\_location | עובד באותה מדינה שגר בה:   * 1. כן * 0. לא | בינארי |
| continent | יבשת:   * Americas * Asia * Europe * Oceania * Africa | קטגוריאלי |

## תיאור והסבר מפורט של המשתנים במחקר הסטטיסטי

**Salary – שכר:** הוא המשתנה המוסבר, משתנה רציף, מתאר את התגמול שמקבל העובד כל חודש בדולרים כלומר השכר החודשי בדולרים. לפי הגרף ניתן לראות שיש זנב ימני זה אומר שיש אוכלוסייה שמקבלת משכורת חודשית גבוהה, אבל אותה אוכלוסייה לא גדולה, לעומת זאת ניתן לראות שהאוכלוסייה של אנשים שמקבלת פחות כסף גדולה יותר. לפי הגרף וההתפלגות של השכר בגרף ניתן להגיד שמדובר בהתפלגות הגיונית, גם במשק ניתן לראות שיש פחות אנשים שמקבלים משכורת גבוהה לעומת אנשים שמקבלים משכורת נמוכה יותר.

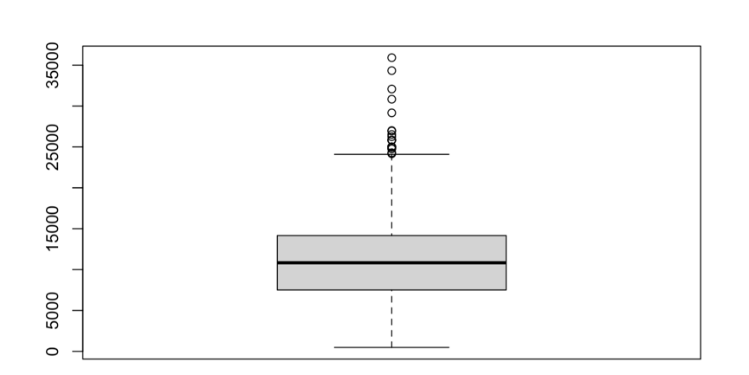
Chart

Description automatically generated

שכר מינימלי בחודש: 473.2 ושכר מקסימאלי בחודש: 35913.9 וכמו כן לפי תרשים הקופסה יש תצפיות חריגות מעל של שכר כמעט 23000

Text

Description automatically generated with low confidence



**Employment Type– סוג משרה:**

לפי התרשים ניתן לראות שרוב הנתונים נאספו ממשרות מלאות, כתוצאה מכך אנחנו בהמשך נתמקד רק בנתונים של משרה מלאה.

Chart, bar chart

Description automatically generated

**Job Title– תפקיד:**

לפי סט הנתונים יש 5 תפקידים שונים מתחום הנתונים: מנתח נתונים, מדען נתונים, מהנדס נתונים, מהנדס מערכות לומדות ומהנדס ניתוח. לפי הגרף ניתן לראות רוב האנשים בעלי תפקיד מהנדס נתונים.

Chart, bar chart, funnel chart

Description automatically generated

**Work year– שנת עבודה:**

המשתנה שנת עבודה מתאר את השכר החודשי של העובד לפי השנה שעבד בה. כלומר כמה העובד מרווח לפי שנה מסוימת. לפי הגרף ניתן לראות שהעובדים דיווחו על משכורת חודשית גבוהה בשנת 2022. לעומת 2020,2021, 2023.

Chart, bar chart

Description automatically generated

**Experience Level– רמת ניסיון (וותק):**

וותק בעבודה הוא משתנה קטגוריאלי, נבדק כמה עובדים יש לפי הוותק בעבודה. לפי הגרף והדיווחים ניתן לראות שהמספר של העובדים הכי גבוהה הוא אנשים שעובדים כ- seniors.

**Chart, bar chart

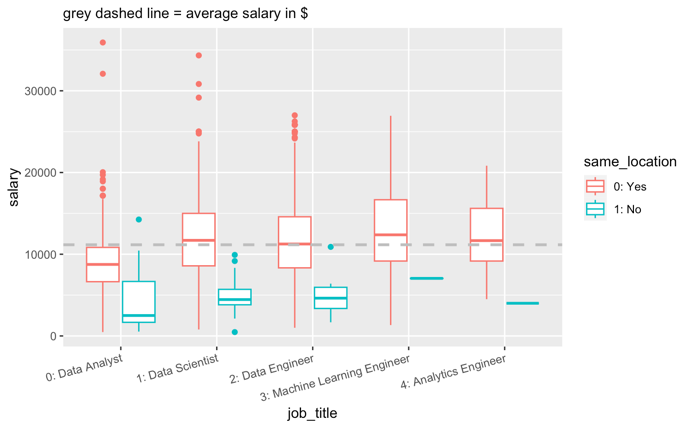
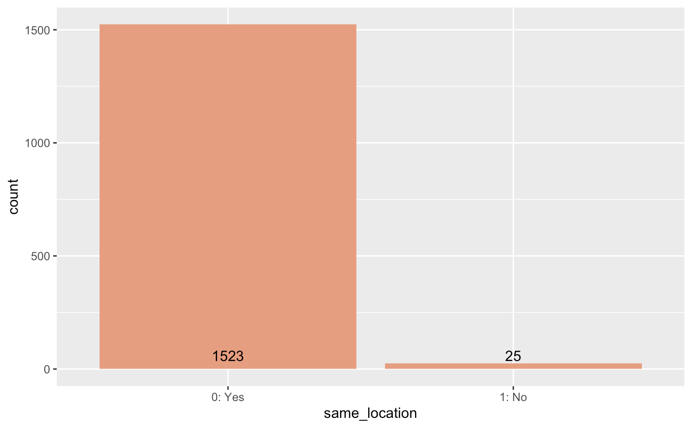
Description automatically generatedChart, box and whisker chart

Description automatically generated**

בנוסף לכך, בתרשים של קופסאות רואים שאין כמעט אנשים ברמה ניהולית שעובדים בתפקידי מהנדסי מערכות לומדות או מנתחי נתונים, והשכר ממוצע מעל 10,000 דולר.

**Same Location–עובד באותה מדינה שגר בה:**

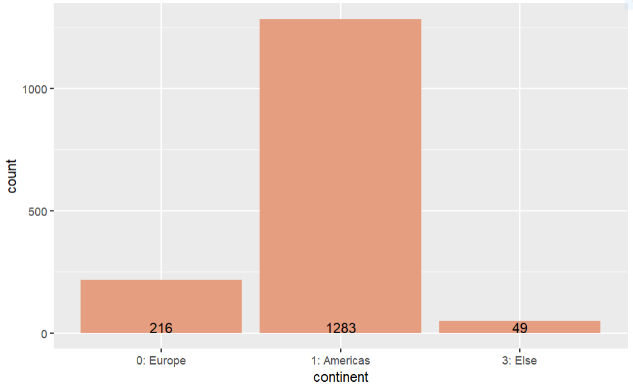
משתנה בינארי שמתאר אם העובד עובד באותו מדינה שהוא גר בה ו98.38% מסט הנתונים שלנו הם עובדים שעובדים באותו מדינה שגרים בה

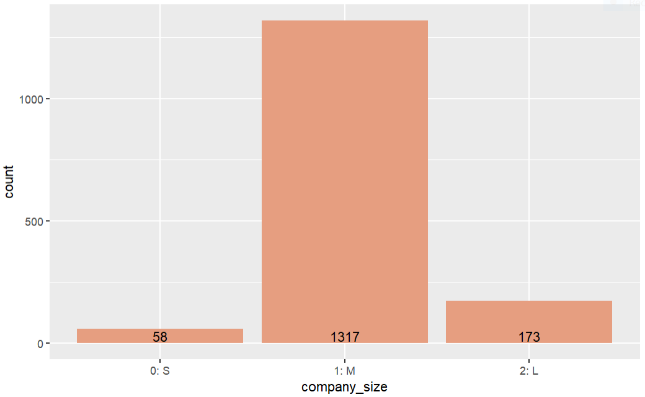


יש הבדל בולט בשכר עבור כל התפקידים בין אם העובד עובד באותו מדינה שהוא גר בה או במדינה אחרת. ננסה לבדוק בהמשך אם יש אינטראקציה בין שני המשתנים same\_location ן- job\_title.

**Continent–יבשת:**

משתנה קטגוריאלי שמתאר את יבשת של מיקום העבודה, בסט הנתונים יש לנו נתונים מארבע יבשות חל ק גדול מהם מאמריקה וחלק ממש זניח מאסיה ואוקיאניה לכן חלטנו למזג אותם לאסיה.Chart, bar chart

Description automatically generated

**Company\_Size–גודל חברה:**

רוב העובדים התחום עובדים בחברות בינוניות.

מ-Box-Plot, רואים שחברות בינוניות נותנים שכר יותר גבוה מאשר חברות גדולות.

היה גם מוזר לראות שבתפקיד-Analytics Engineer לא עובדים בחברות גדולות. אנו חושבים שבחברות גדולות נותנים שם אחר לתפקיד זה.

Chart, bar chart

Description automatically generatedChart, box and whisker chart

Description automatically generated

לפני שנתקדם לשלב הבא – בניית מודלים, החלטנו לוותר על חלק מהתצפיות שנראים לנו חריגות. תצפיות אלו הינן שכר חודשי קטן מ-1000 דולר בחודש. אנו חושבים שהתצפיות אלו לא היגיוניות ומצב כזה לא קורה במציאות, ולכן, נמשיך בשלב הבא בלי תצפיות אלו.

## מודלים וניתוח סטטיסטי

כדי לחזות את השכר החלטנו להשתמש שני מודלים שהם רגרסיה ליניארית ויער אקראי.

1. רגרסיה ליניארית

המודל הראשון שהחלטנו למדל בו את הנתונים הוא מודל של רגרסיה ליניארית. מודל זה יעזור לנו לחזות את השכר, וגם כן להבין את ההשפעה של המשתנים הבלתי תלויים על השכר.

המודל ההתחלתי שלנו יכלול את המשתנים הבאים: job\_title\*(experience\_level + company\_size + same\_location + work\_year + remote\_ratio + continent), רוצים לבדוק גם את האינטראקציה בין תפקידים שונים עם שאר המשתנים הבלתי תלויים.

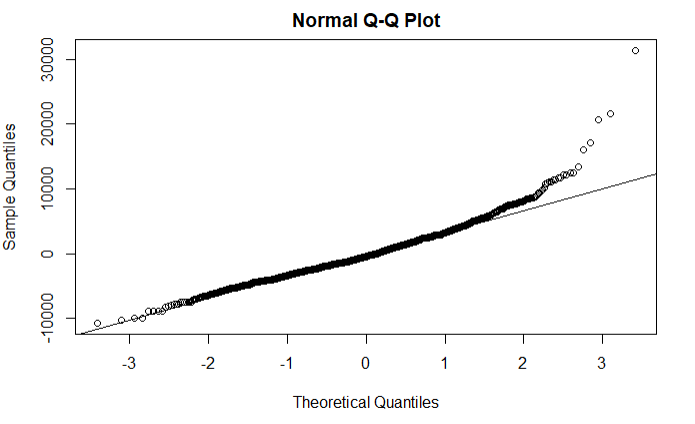
מכיוון שיש הרבה משתנים במודל אם רוצים גם לראות את השפעת האינטראקציה, נתחיל את הניתוח שלנו ברגרסיה בצעדים(both). הבדיקה נעשתה בעזרת המדד-AIC. המודל שהתקבל הינו:

הנחות המודל

* הנחת הנורמליות
* שוויון שונויות
* ליניאריות
* טעויות בלתי תלויות

לפני שנסביר את תוצאות המודל, נבדוק את ההנחות:

1. הנחת הנורמליות:



Shapiro-Wilk normality test

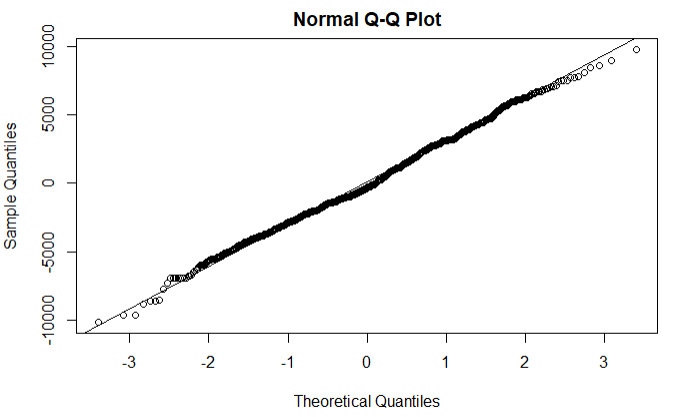
data: residuals

W = 0.95216, p-value < 2.2e-16

רואים שיש בריחה בזנב עליון, הגם לפי מבחן SW, ההנחה של הנורמליות נדחית.

ננסה לשפר ולפתור את הבעיה הזאת. נריץ עכשיו את המודל החדש ברגרסיה בצעדים עבור שכר הקטן מ- 20000 דולר בחודש. המודל המתקבל:

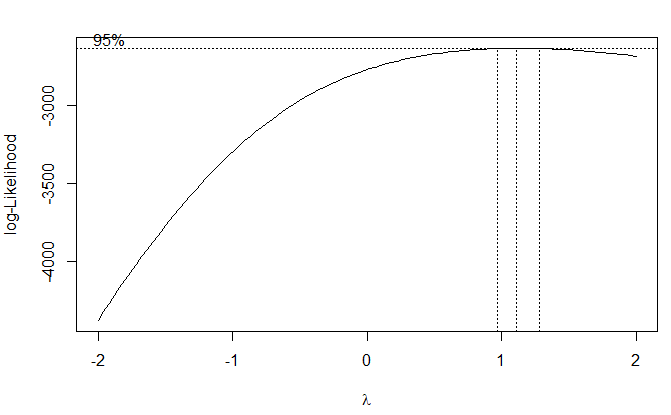
Chart

Description automatically generatedבודקים את ההנחה מחדש:

Shapiro-Wilk normality test

data: residuals

W = 0.9957, p-value = 0.0003502

Chart, line chart

Description automatically generatedלמרות שתרשים QQPLOT נראה טוב, ההשערה של הנחת הנורמליות נדחית, ויש צורך בהתמרה. רואים שהתמרת השורש הינה בתוך הרווח סמך ב-Box-Cox. ולכן, נחזור להתאמת מודל חדש על התמרת שורת עבור השכר נריץ רגרסיה בצעדים ונקבל:

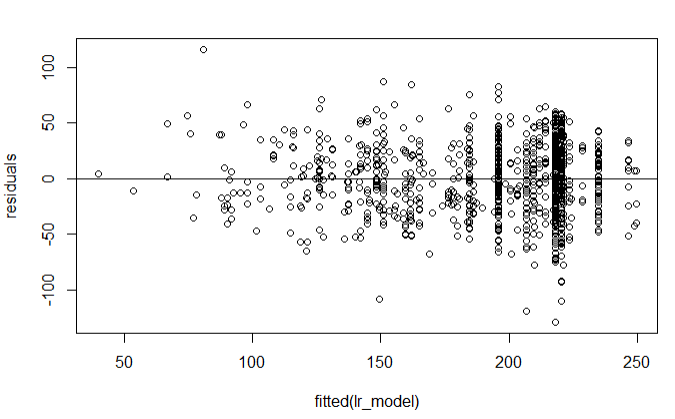
Shapiro-Wilk normality test :data: residuals

W = 0.99447, p-value = 2.909e-05

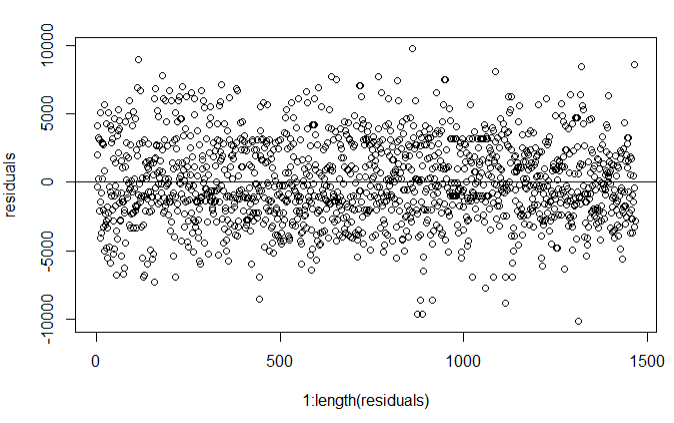
לפי Box-Cox, רואים שעכשיו נמצאת בתוך הרווח סמך.

לפי QQPLOT, אפשר לראות שבאמצע המצב השתפר, אבל יש בריחה בזנבות שהיא בולטת יותר.

מבחן Shapiro-Wilks גם אחרי ההתמרה דחה את ההשערה של הנורמליות.

1. הנחת שוויון שונויות:

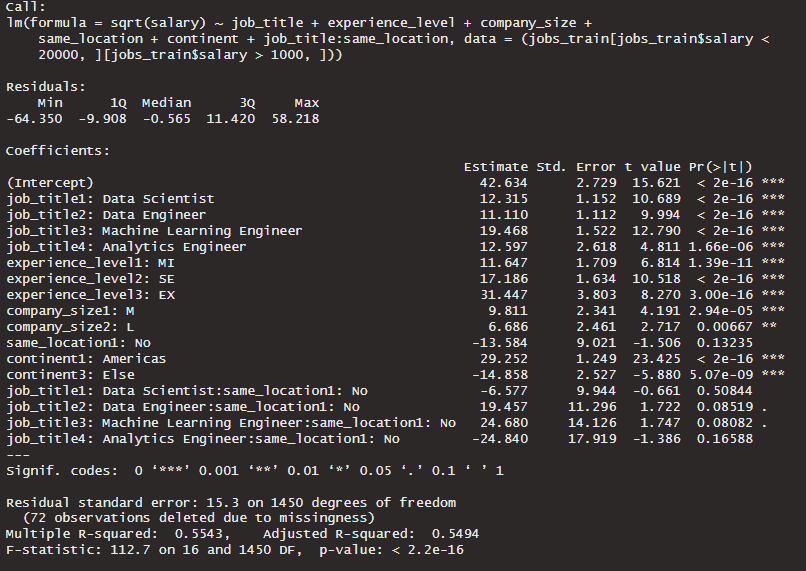
מהתרשים, אפשר לראות שהפיזור לא נראה הכי טוב, למרות שאין פיזור של "קונוס", רואים שהרבה שאריות נופלות באזור של +200~. כלומר ההנחה לא תתקיים.

1. שאריות בלתי תלויות

לפי התרשים לא רואים סוג של מגמתיות בנתונים, אפשר לראות שההנחה כן תתקיים.

אחרי ההתמרה של משתנה המטרה בהתמרת השורש, ראינו ששתי ההנחות, נורמליות ושוויון שונויות לא מתקיימות. ולכן המודל לא נמצא "במצב טוב" למרות הניסיון בהתמרה. המשתנה שכר גם כן, אינו משתנה של ספירה, זמן או משתנה חסום, ולכן לא ננסה להשתמש במודלים מוכללים או התמרות של לוג.

למרות אי-ההתקיימות של הנחות המודל, אנחנו נציג את התוצאות שלו, אבל לא נסמך על התוצאות שייתנו מידע נכון ואמיתי.



מסקנות מהפלט:

הסבר האומדנים:

* החותך הינו דאטה אנליסט מתחיל בחברה קטנה באותה מדינה שהוא גר בה. השכר לנו יהיה שווה ל- דולר בחודש בממוצע.
  + השינוי בתפקיד יעלה את השכר החודשי ב-:
    - Data Scientist: דולר בחודש בממוצע.
    - Data Engineer: דולר בחודש בממוצע.
    - ML Engineer: דולר בחודש בממוצע.
    - Analytics Engineer: דולר בחודש בממוצע.
  + השינוי של גודל החברה:
    - מעלה את השכר ב-96.26 דולר בממוצע במעבר בחברה בגודל בינוני.
    - מעלה את השכר ב-44.7 דולר בממוצע במעבר בחברה גדולה.
  + השינוי ברמת התפקיד הינו חיובי
    - 135.66+ דולר בתפקיד MI
    - 295.36+ דולר בתפקיד Senior
    - 988.91+ דולר בתפקיד Expert
  + השכר יורד ב- 184.41 דולר בממוצע כשעובד בחול מאשר במדינה שנמצא בה העובד. התוצאה לא מובהקת סטטיסטית.
  + באמריקה או קנדה השכר שלהם גבוה ב- 855.68 דולר בממוצע מאשר באירופה, ונמוך ב-220.77 דולר בחודש בממוצע ביבשת אחרת.
* בתפקיד של Data Scientist ו- Analytics Engineer יורד אם עובדים במדינה אחרת ועולה בתפקידי Data/ML Engineer. תוצאות אלו אינות מובהקות סטטיסטית.
* ה-
* המודל מובהק

1. יער אקראי – random forest.

ממודל רגרסיה ליניארית ראינו שהנחת הנורמליות הלא התקבלה באופן סטטיסטי, ויכול להיות שתוצאות המודל יכולים להשתפר. ולכן, החלטנו להשתמש במודל נוסף: יער אקראי.

Random Forest הוא שיטת למידה אנסמבלית המשלבת עצי החלטה מרובים ליצירת מודל חיזוי מדויק וחזק יותר. היתרון של המודל הנוכחי בהשוואה עם רגרסיה ליניארית הוא שבמודל זה אין הנחות. אבל החיסרון הבולט שלו הוא הקושי להסביר את תוצאותיו.

ביער אקאי יש פרמטרים שצריך לבדוק לפני. אלה פרמטרים של המודל שאינם נלמדים במהלך האימון, אך יש להגדיר אותם לפני האימון. ההיפרפרמטרים במודל שבדנו:

* mtry: מספר התכונות שיש לקחת בחשבון בעת פיצול כל צומת.
* min.node.size: המספר המינימלי של דגימות הנדרש לפיצול צומת.
* splitrule: כלל הפיצול לשימוש בכל צומת:
  + Extratrees: הסף עבור כל תכונה נבחר באופן אקראי, במקום להתבסס על הערכים האמיתיים של התכונה בנתוני האימון. זה יכול לעזור להפחית התאמה יתר ולשפר את ביצועי ההכללה של המודל.
  + שונות (הידועה גם בשם "mse" או שגיאה בריבוע ממוצעת): כלל פיצול זה בוחר את התכונה והסף הממזערים את השונות של משתנה היעד (כלומר, סכום ההבדלים בריבוע בין הערכים האמיתיים והחזויים של משתנה היעד). כלל הפיצול הזה מתאים לבעיות רגרסיה.

בדקנו את ההיפרפרמטרים האלה בעזרת שיטת K-Folds Cross Validation כך ש-K=5. תוצאות הבדיקה מוצגים בגרף הבא:

## 

ההיפרפרמטרים שעזרו להגיע למינימום RMSE הם: mtry = 15, target.node.size = 20 ו- splitrule = extratrees.

### השוואת שני המודלים

נשווה את תוצאות שני המודלים לפי קריטריון RMSE עבור test set שיש לנו. חשוב לציין שקודם כל צריך להשוות אותם כאשר הם באותה סקאלה. מכיוון שבמודל רגרסיה ליניארית עשיני התמרת שורש לשכר, אנחנו נחזה את השכר בשורש ונפעיל את הפעולה ההפוכה לשורש, שהיא החזקה, ואחר כך נחשב את RMSE למודלים ונשווה אותם.

|  |  |  |
| --- | --- | --- |
| Linear regression | Random forest | test |
| 12098.584 | 13173.02 | 10775 |
| 9778.041 | 10063.56 | 15491.67 |
| 14007.087 | 14559.39 | 12783.33 |
| 25833.333 | 17594.69 | 15439.189 |
| 14350.000 | 14559.39 | 14007.087 |

אחרי חישוב RMSE, נציג את 5 תחזיות ואת השכר עבורן, ואת RMSE של שני המודלים:

יש הבדל נמוך בין שניהם. אבל אנו נעדיף להשתמש ב-Random Forest, מכיוון שעבורו ה-RMSE נמוך יותר, וגם בגלל שבמודל רגרסיה ליניארית יש הנחות שלא התקיימו.

# דיון ומסקנות

עבודת מחקר זו בוחנת את הגורמים המשפיעים על שכר העבודה של תפקדים מבוססים נתונים כמו ניסיון, גודל חברה, היקף משרה ומיקום. המחקר מספק תובנות חשובות לגבי האופי המורכב של שכר העבודה בנוסף לכך ראינו שיהיה לנו משתנים בכלל לא מאוזנים או ייצוג גדולים יותר מאחרים, מה שמוביל לפיזור לא אחיד של נתונים ולהטיות פוטנציאליות בניתוח או בקבלת החלטות טופלו כדי שנגיע לחלטה כמה שיותר נכונה.

כדי לחזות את השכר, השתמשנו בשני מודלים שהם רגרסיה ליניארית עם התמרת שורש למשתנה המטרה. הנחות המודל לא התקיימו כמו שציפינו. ראינו שעבור המודל, מתוצאות המודל, ML Engineer שגר באמריקה הוא המרוויח ביותר, השכר שלו עולה גבוה יותר כשעובר בחברה בינונית. השכר גם כן עולה כשהוא מתקדם בתפקיד שלו. המודל השני שהשתמשנו בו הוא יער אקראי מסיבה שמודל זה ללא הנחות וגם תוצאות מודל הרגרסיה הליניארית לא נוכל לסמוך עליהם.

אחרי בניית המודלים, השווינו את התחזיות של שני המודלים על סט נתונים חדש, וראינו שמודל random forest מזער את מדד RMSE יותר מאשר מודל הרגרסיה הליניארית.

למרות שהמחקר מספק תובנות חשובות לגבי השכר של משרות מבוססים נתונים, חשוב להכיר בכך שהמחקר מוגבל לנתוני שכר שנתי ואינו מתחשב בגורמים כמו מגדר, וגיל. לכן, חיוני לפרש את הממצאים בזהירות.

לפרויקטים הבאים, נמליץ לבעלי האתר לאסוף עוד נתונים על רמת תואר, שנות ניסיות בתפקיד, גיל, מין, ואת התואר( תואר בהנדסה, מדעים...). מעניין לנו לראות איך הם משפיעים על השכר, ואנחנו חושבים שמידע זה יכול לשפר את התחזיות של המודלים שהצגנו.

נספחים

מודל רגרסיה ליניארית 1

lr\_model <- MASS::stepAIC(lm((salary) ~ job\_title\*(experience\_level + company\_size + same\_location + work\_year + remote\_ratio + continent) ,  
 data = jobs\_train), direction = "both")

## Start: AIC=25555.74  
## (salary) ~ job\_title \* (experience\_level + company\_size + same\_location +   
## work\_year + remote\_ratio + continent)  
##   
## Df Sum of Sq RSS AIC  
## - job\_title:work\_year 10 61992308 2.1182e+10 25540  
## - job\_title:company\_size 6 26761749 2.1147e+10 25546  
## - job\_title:continent 7 77679673 2.1198e+10 25547  
## - job\_title:remote\_ratio 7 94280943 2.1214e+10 25549  
## - job\_title:same\_location 3 60479184 2.1181e+10 25554  
## - job\_title:experience\_level 10 269791027 2.1390e+10 25555  
## <none> 2.1120e+10 25556  
##   
## Step: AIC=25540.28  
## (salary) ~ job\_title + experience\_level + company\_size + same\_location +   
## work\_year + remote\_ratio + continent + job\_title:experience\_level +   
## job\_title:company\_size + job\_title:same\_location + job\_title:remote\_ratio +   
## job\_title:continent  
##   
## Df Sum of Sq RSS AIC  
## - job\_title:company\_size 6 21944380 2.1204e+10 25530  
## - job\_title:continent 7 76532613 2.1259e+10 25532  
## - job\_title:remote\_ratio 7 89029998 2.1271e+10 25533  
## - job\_title:same\_location 3 55484623 2.1238e+10 25538  
## <none> 2.1182e+10 25540  
## - job\_title:experience\_level 10 276421393 2.1459e+10 25540  
## - work\_year 3 160812616 2.1343e+10 25546  
## + job\_title:work\_year 10 61992308 2.1120e+10 25556  
##   
## Step: AIC=25529.88  
## (salary) ~ job\_title + experience\_level + company\_size + same\_location +   
## work\_year + remote\_ratio + continent + job\_title:experience\_level +   
## job\_title:same\_location + job\_title:remote\_ratio + job\_title:continent  
##   
## Df Sum of Sq RSS AIC  
## - job\_title:continent 7 76027456 2.1280e+10 25521  
## - job\_title:remote\_ratio 7 94788230 2.1299e+10 25523  
## - job\_title:same\_location 4 84307711 2.1288e+10 25528  
## <none> 2.1204e+10 25530  
## - job\_title:experience\_level 10 278030675 2.1482e+10 25530  
## - company\_size 2 114858619 2.1319e+10 25534  
## - work\_year 3 158830403 2.1363e+10 25535  
## + job\_title:company\_size 6 21944380 2.1182e+10 25540  
## + job\_title:work\_year 10 57174939 2.1147e+10 25546  
##   
## Step: AIC=25521.42  
## (salary) ~ job\_title + experience\_level + company\_size + same\_location +   
## work\_year + remote\_ratio + continent + job\_title:experience\_level +   
## job\_title:same\_location + job\_title:remote\_ratio  
##   
## Df Sum of Sq RSS AIC  
## - job\_title:remote\_ratio 7 95507407 2.1376e+10 25514  
## - job\_title:same\_location 4 80686064 2.1361e+10 25519  
## <none> 2.1280e+10 25521  
## - job\_title:experience\_level 10 318291840 2.1598e+10 25524  
## - company\_size 2 107481363 2.1388e+10 25525  
## - work\_year 3 163876763 2.1444e+10 25527  
## + job\_title:continent 7 76027456 2.1204e+10 25530  
## + job\_title:company\_size 6 21439222 2.1259e+10 25532  
## + job\_title:work\_year 10 54895697 2.1225e+10 25537  
## - continent 2 4535281870 2.5815e+10 25817  
##   
## Step: AIC=25514.35  
## (salary) ~ job\_title + experience\_level + company\_size + same\_location +   
## work\_year + remote\_ratio + continent + job\_title:experience\_level +   
## job\_title:same\_location  
##   
## Df Sum of Sq RSS AIC  
## - remote\_ratio 2 14486091 2.1390e+10 25511  
## - job\_title:same\_location 4 87694682 2.1463e+10 25513  
## <none> 2.1376e+10 25514  
## - company\_size 2 112995639 2.1489e+10 25519  
## - work\_year 3 168746148 2.1544e+10 25521  
## + job\_title:remote\_ratio 7 95507407 2.1280e+10 25521  
## - job\_title:experience\_level 10 393663454 2.1769e+10 25523  
## + job\_title:continent 7 76746632 2.1299e+10 25523  
## + job\_title:company\_size 6 21992723 2.1354e+10 25525  
## + job\_title:work\_year 10 47246115 2.1328e+10 25531  
## - continent 2 4535239317 2.5911e+10 25808  
##   
## Step: AIC=25511.4  
## (salary) ~ job\_title + experience\_level + company\_size + same\_location +   
## work\_year + continent + job\_title:experience\_level + job\_title:same\_location  
##   
## Df Sum of Sq RSS AIC  
## - job\_title:same\_location 4 84676657 2.1475e+10 25510  
## <none> 2.1390e+10 25511  
## + remote\_ratio 2 14486091 2.1376e+10 25514  
## - company\_size 2 110144444 2.1500e+10 25515  
## - work\_year 3 162641314 2.1553e+10 25517  
## - job\_title:experience\_level 10 387968274 2.1778e+10 25519  
## + job\_title:continent 7 75153711 2.1315e+10 25520  
## + job\_title:company\_size 6 24002047 2.1366e+10 25522  
## + job\_title:work\_year 10 48254917 2.1342e+10 25528  
## - continent 2 4730514926 2.6121e+10 25817  
##   
## Step: AIC=25509.52  
## (salary) ~ job\_title + experience\_level + company\_size + same\_location +   
## work\_year + continent + job\_title:experience\_level  
##   
## Df Sum of Sq RSS AIC  
## <none> 2.1475e+10 25510  
## + job\_title:same\_location 4 84676657 2.1390e+10 25511  
## + remote\_ratio 2 11468066 2.1463e+10 25513  
## - company\_size 2 124588841 2.1599e+10 25515  
## - work\_year 3 156663378 2.1631e+10 25515  
## - same\_location 1 120525700 2.1595e+10 25516  
## - job\_title:experience\_level 10 386668124 2.1861e+10 25517  
## + job\_title:continent 7 70251486 2.1405e+10 25518  
## + job\_title:company\_size 7 64771211 2.1410e+10 25519  
## + job\_title:work\_year 10 48152092 2.1427e+10 25526  
## - continent 2 4681035885 2.6156e+10 25811

summary(lr\_model)

##   
## Call:  
## lm(formula = (salary) ~ job\_title + experience\_level + company\_size +   
## same\_location + work\_year + continent + job\_title:experience\_level,   
## data = jobs\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10682.9 -2397.8 -454.5 2152.4 31424.4   
##   
## Coefficients: (2 not defined because of singularities)  
## Estimate Std. Error  
## (Intercept) 897.6 1026.2  
## job\_title1: Data Scientist 2073.3 939.9  
## job\_title2: Data Engineer 3430.6 955.5  
## job\_title3: Machine Learning Engineer 4776.4 1217.1  
## job\_title4: Analytics Engineer 3666.7 3429.0  
## experience\_level1: MI 2860.1 818.9  
## experience\_level2: SE 3154.0 775.2  
## experience\_level3: EX 3291.2 2755.7  
## company\_size1: M 1760.1 604.2  
## company\_size2: L 1246.7 601.3  
## same\_location1: No -2322.8 794.8  
## work\_year1: 2021 -1537.4 740.6  
## work\_year2: 2022 -1028.2 703.5  
## work\_year3: 2023 -478.0 726.7  
## continent1: Americas 5079.3 308.6  
## continent3: Else -1490.0 624.0  
## job\_title1: Data Scientist:experience\_level1: MI -729.6 1094.0  
## job\_title2: Data Engineer:experience\_level1: MI -2171.3 1093.4  
## job\_title3: Machine Learning Engineer:experience\_level1: MI -1904.7 1395.2  
## job\_title4: Analytics Engineer:experience\_level1: MI -2854.3 3778.0  
## job\_title1: Data Scientist:experience\_level2: SE 1135.5 1001.2  
## job\_title2: Data Engineer:experience\_level2: SE -527.2 1011.4  
## job\_title3: Machine Learning Engineer:experience\_level2: SE 311.8 1293.8  
## job\_title4: Analytics Engineer:experience\_level2: SE -100.9 3499.5  
## job\_title1: Data Scientist:experience\_level3: EX 5059.9 3559.5  
## job\_title2: Data Engineer:experience\_level3: EX 6079.6 2941.9  
## job\_title3: Machine Learning Engineer:experience\_level3: EX NA NA  
## job\_title4: Analytics Engineer:experience\_level3: EX NA NA  
## t value Pr(>|t|)  
## (Intercept) 0.875 0.381899  
## job\_title1: Data Scientist 2.206 0.027543  
## job\_title2: Data Engineer 3.590 0.000341  
## job\_title3: Machine Learning Engineer 3.924 9.08e-05  
## job\_title4: Analytics Engineer 1.069 0.285098  
## experience\_level1: MI 3.492 0.000492  
## experience\_level2: SE 4.068 4.97e-05  
## experience\_level3: EX 1.194 0.232541  
## company\_size1: M 2.913 0.003630  
## company\_size2: L 2.073 0.038321  
## same\_location1: No -2.923 0.003521  
## work\_year1: 2021 -2.076 0.038070  
## work\_year2: 2022 -1.462 0.144064  
## work\_year3: 2023 -0.658 0.510843  
## continent1: Americas 16.459 < 2e-16  
## continent3: Else -2.388 0.017069  
## job\_title1: Data Scientist:experience\_level1: MI -0.667 0.504971  
## job\_title2: Data Engineer:experience\_level1: MI -1.986 0.047235  
## job\_title3: Machine Learning Engineer:experience\_level1: MI -1.365 0.172397  
## job\_title4: Analytics Engineer:experience\_level1: MI -0.756 0.450053  
## job\_title1: Data Scientist:experience\_level2: SE 1.134 0.256940  
## job\_title2: Data Engineer:experience\_level2: SE -0.521 0.602279  
## job\_title3: Machine Learning Engineer:experience\_level2: SE 0.241 0.809575  
## job\_title4: Analytics Engineer:experience\_level2: SE -0.029 0.977008  
## job\_title1: Data Scientist:experience\_level3: EX 1.422 0.155366  
## job\_title2: Data Engineer:experience\_level3: EX 2.067 0.038943  
## job\_title3: Machine Learning Engineer:experience\_level3: EX NA NA  
## job\_title4: Analytics Engineer:experience\_level3: EX NA NA  
##   
## (Intercept)   
## job\_title1: Data Scientist \*   
## job\_title2: Data Engineer \*\*\*  
## job\_title3: Machine Learning Engineer \*\*\*  
## job\_title4: Analytics Engineer   
## experience\_level1: MI \*\*\*  
## experience\_level2: SE \*\*\*  
## experience\_level3: EX   
## company\_size1: M \*\*   
## company\_size2: L \*   
## same\_location1: No \*\*   
## work\_year1: 2021 \*   
## work\_year2: 2022   
## work\_year3: 2023   
## continent1: Americas \*\*\*  
## continent3: Else \*   
## job\_title1: Data Scientist:experience\_level1: MI   
## job\_title2: Data Engineer:experience\_level1: MI \*   
## job\_title3: Machine Learning Engineer:experience\_level1: MI   
## job\_title4: Analytics Engineer:experience\_level1: MI   
## job\_title1: Data Scientist:experience\_level2: SE   
## job\_title2: Data Engineer:experience\_level2: SE   
## job\_title3: Machine Learning Engineer:experience\_level2: SE   
## job\_title4: Analytics Engineer:experience\_level2: SE   
## job\_title1: Data Scientist:experience\_level3: EX   
## job\_title2: Data Engineer:experience\_level3: EX \*   
## job\_title3: Machine Learning Engineer:experience\_level3: EX   
## job\_title4: Analytics Engineer:experience\_level3: EX   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3756 on 1522 degrees of freedom  
## Multiple R-squared: 0.4479, Adjusted R-squared: 0.4388   
## F-statistic: 49.39 on 25 and 1522 DF, p-value: < 2.2e-16

###check assumptions

1. normality assumptions:

# Extract the residuals  
residuals <- residuals(lr\_model)  
  
# Plot the Q-Q plot of the residuals  
qqnorm(residuals)  
qqline(residuals)

Chart, line chart

Description automatically generated

shapiro.test(residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: residuals  
## W = 0.95216, p-value < 2.2e-16

we will try to make it better: first we will remove outliers: 20k$+ per month:

מודל רגרסיה ליניארית 2  
lr\_model <- MASS::stepAIC(lm((salary) ~ job\_title\*(experience\_level + company\_size + same\_location + work\_year + remote\_ratio + continent) ,  
 data = jobs\_train[jobs\_train$salary<20000,][jobs\_train$salary>1000,]), direction = "both")

## Start: AIC=23606.75  
## (salary) ~ job\_title \* (experience\_level + company\_size + same\_location +   
## work\_year + remote\_ratio + continent)  
##   
## Df Sum of Sq RSS AIC  
## - job\_title:work\_year 10 64164379 1.3196e+10 23594  
## - job\_title:continent 7 44848525 1.3177e+10 23598  
## - job\_title:company\_size 6 29181037 1.3161e+10 23598  
## - job\_title:remote\_ratio 7 47775013 1.3180e+10 23598  
## - job\_title:experience\_level 10 137182225 1.3269e+10 23602  
## <none> 1.3132e+10 23607  
## - job\_title:same\_location 3 62587177 1.3194e+10 23608  
##   
## Step: AIC=23593.9  
## (salary) ~ job\_title + experience\_level + company\_size + same\_location +   
## work\_year + remote\_ratio + continent + job\_title:experience\_level +   
## job\_title:company\_size + job\_title:same\_location + job\_title:remote\_ratio +   
## job\_title:continent  
##   
## Df Sum of Sq RSS AIC  
## - job\_title:company\_size 6 22899481 1.3219e+10 23584  
## - job\_title:continent 7 42675090 1.3239e+10 23585  
## - job\_title:remote\_ratio 7 53267961 1.3249e+10 23586  
## - job\_title:experience\_level 10 148756076 1.3345e+10 23590  
## - work\_year 3 41010247 1.3237e+10 23593  
## <none> 1.3196e+10 23594  
## - job\_title:same\_location 3 63128645 1.3259e+10 23595  
## + job\_title:work\_year 10 64164379 1.3132e+10 23607  
##   
## Step: AIC=23584.44  
## (salary) ~ job\_title + experience\_level + company\_size + same\_location +   
## work\_year + remote\_ratio + continent + job\_title:experience\_level +   
## job\_title:same\_location + job\_title:remote\_ratio + job\_title:continent  
##   
## Df Sum of Sq RSS AIC  
## - job\_title:continent 7 41891289 1.3261e+10 23575  
## - job\_title:remote\_ratio 7 61383200 1.3280e+10 23577  
## - job\_title:experience\_level 10 153321317 1.3372e+10 23581  
## - work\_year 3 38732474 1.3258e+10 23583  
## <none> 1.3219e+10 23584  
## - job\_title:same\_location 4 95315645 1.3314e+10 23587  
## - company\_size 2 92308634 1.3311e+10 23591  
## + job\_title:company\_size 6 22899481 1.3196e+10 23594  
## + job\_title:work\_year 10 57882823 1.3161e+10 23598  
##   
## Step: AIC=23575.09  
## (salary) ~ job\_title + experience\_level + company\_size + same\_location +   
## work\_year + remote\_ratio + continent + job\_title:experience\_level +   
## job\_title:same\_location + job\_title:remote\_ratio  
##   
## Df Sum of Sq RSS AIC  
## - job\_title:remote\_ratio 7 62627984 1.3323e+10 23568  
## - job\_title:experience\_level 10 155752674 1.3417e+10 23572  
## - work\_year 3 43889498 1.3305e+10 23574  
## <none> 1.3261e+10 23575  
## - job\_title:same\_location 4 95723907 1.3356e+10 23578  
## - company\_size 2 81301034 1.3342e+10 23580  
## + job\_title:continent 7 41891289 1.3219e+10 23584  
## + job\_title:company\_size 6 22115680 1.3239e+10 23585  
## + job\_title:work\_year 10 54445092 1.3206e+10 23589  
## - continent 2 4355660939 1.7616e+10 23988  
##   
## Step: AIC=23568  
## (salary) ~ job\_title + experience\_level + company\_size + same\_location +   
## work\_year + remote\_ratio + continent + job\_title:experience\_level +   
## job\_title:same\_location  
##   
## Df Sum of Sq RSS AIC  
## - remote\_ratio 2 16736096 1.3340e+10 23566  
## - work\_year 3 44912803 1.3368e+10 23567  
## <none> 1.3323e+10 23568  
## - job\_title:experience\_level 10 185062673 1.3508e+10 23568  
## - job\_title:same\_location 4 107184417 1.3431e+10 23572  
## - company\_size 2 86015084 1.3409e+10 23573  
## + job\_title:remote\_ratio 7 62627984 1.3261e+10 23575  
## + job\_title:company\_size 6 28782353 1.3295e+10 23577  
## + job\_title:continent 7 43136073 1.3280e+10 23577  
## + job\_title:work\_year 10 62256932 1.3261e+10 23581  
## - continent 2 4324739087 1.7648e+10 23976  
##   
## Step: AIC=23565.84  
## (salary) ~ job\_title + experience\_level + company\_size + same\_location +   
## work\_year + continent + job\_title:experience\_level + job\_title:same\_location  
##   
## Df Sum of Sq RSS AIC  
## - work\_year 3 34339405 1.3374e+10 23564  
## - job\_title:experience\_level 10 177195560 1.3517e+10 23565  
## <none> 1.3340e+10 23566  
## + remote\_ratio 2 16736096 1.3323e+10 23568  
## - job\_title:same\_location 4 106351298 1.3446e+10 23570  
## - company\_size 2 82515667 1.3423e+10 23571  
## + job\_title:company\_size 6 29115767 1.3311e+10 23575  
## + job\_title:continent 7 42187085 1.3298e+10 23575  
## + job\_title:work\_year 10 63548307 1.3277e+10 23579  
## - continent 2 4507159572 1.7847e+10 23989  
##   
## Step: AIC=23563.61  
## (salary) ~ job\_title + experience\_level + company\_size + same\_location +   
## continent + job\_title:experience\_level + job\_title:same\_location  
##   
## Df Sum of Sq RSS AIC  
## <none> 1.3374e+10 23564  
## - job\_title:experience\_level 10 187354939 1.3562e+10 23564  
## + work\_year 3 34339405 1.3340e+10 23566  
## - job\_title:same\_location 4 100360333 1.3475e+10 23567  
## + remote\_ratio 2 6162698 1.3368e+10 23567  
## + job\_title:continent 7 46844678 1.3328e+10 23573  
## + job\_title:company\_size 6 24645665 1.3350e+10 23573  
## - company\_size 2 137229985 1.3512e+10 23575  
## - continent 2 4727336544 1.8102e+10 24004

summary(lr\_model)

##   
## Call:  
## lm(formula = (salary) ~ job\_title + experience\_level + company\_size +   
## same\_location + continent + job\_title:experience\_level +   
## job\_title:same\_location, data = jobs\_train[jobs\_train$salary <   
## 20000, ][jobs\_train$salary > 1000, ])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10125.1 -1997.8 -332.1 2178.4 9802.2   
##   
## Coefficients: (2 not defined because of singularities)  
## Estimate Std. Error  
## (Intercept) -114.5 745.4  
## job\_title1: Data Scientist 2095.5 779.1  
## job\_title2: Data Engineer 3356.2 789.3  
## job\_title3: Machine Learning Engineer 3721.8 1032.9  
## job\_title4: Analytics Engineer 3666.7 2782.1  
## experience\_level1: MI 2659.5 674.7  
## experience\_level2: SE 3158.5 637.1  
## experience\_level3: EX 3244.1 2237.5  
## company\_size1: M 1732.9 476.3  
## company\_size2: L 1207.4 498.2  
## same\_location1: No -1074.0 1825.7  
## continent1: Americas 5137.5 252.0  
## continent3: Else -1378.8 506.2  
## job\_title1: Data Scientist:experience\_level1: MI -462.3 900.1  
## job\_title2: Data Engineer:experience\_level1: MI -1925.6 898.9  
## job\_title3: Machine Learning Engineer:experience\_level1: MI -573.1 1163.9  
## job\_title4: Analytics Engineer:experience\_level1: MI -2681.5 3064.7  
## job\_title1: Data Scientist:experience\_level2: SE 758.9 826.9  
## job\_title2: Data Engineer:experience\_level2: SE -1062.3 833.6  
## job\_title3: Machine Learning Engineer:experience\_level2: SE 301.5 1096.6  
## job\_title4: Analytics Engineer:experience\_level2: SE -585.0 2845.7  
## job\_title1: Data Scientist:experience\_level3: EX 1654.5 3145.6  
## job\_title2: Data Engineer:experience\_level3: EX 3642.0 2445.7  
## job\_title3: Machine Learning Engineer:experience\_level3: EX NA NA  
## job\_title4: Analytics Engineer:experience\_level3: EX NA NA  
## job\_title1: Data Scientist:same\_location1: No -2162.7 2013.3  
## job\_title2: Data Engineer:same\_location1: No 1555.2 2285.0  
## job\_title3: Machine Learning Engineer:same\_location1: No 2601.0 2867.0  
## job\_title4: Analytics Engineer:same\_location1: No -6189.2 3594.3  
## t value Pr(>|t|)  
## (Intercept) -0.154 0.877883  
## job\_title1: Data Scientist 2.690 0.007235  
## job\_title2: Data Engineer 4.252 2.25e-05  
## job\_title3: Machine Learning Engineer 3.603 0.000325  
## job\_title4: Analytics Engineer 1.318 0.187724  
## experience\_level1: MI 3.942 8.47e-05  
## experience\_level2: SE 4.957 7.99e-07  
## experience\_level3: EX 1.450 0.147303  
## company\_size1: M 3.638 0.000284  
## company\_size2: L 2.423 0.015504  
## same\_location1: No -0.588 0.556444  
## continent1: Americas 20.388 < 2e-16  
## continent3: Else -2.724 0.006534  
## job\_title1: Data Scientist:experience\_level1: MI -0.514 0.607591  
## job\_title2: Data Engineer:experience\_level1: MI -2.142 0.032355  
## job\_title3: Machine Learning Engineer:experience\_level1: MI -0.492 0.622493  
## job\_title4: Analytics Engineer:experience\_level1: MI -0.875 0.381738  
## job\_title1: Data Scientist:experience\_level2: SE 0.918 0.358861  
## job\_title2: Data Engineer:experience\_level2: SE -1.274 0.202780  
## job\_title3: Machine Learning Engineer:experience\_level2: SE 0.275 0.783431  
## job\_title4: Analytics Engineer:experience\_level2: SE -0.206 0.837143  
## job\_title1: Data Scientist:experience\_level3: EX 0.526 0.598995  
## job\_title2: Data Engineer:experience\_level3: EX 1.489 0.136670  
## job\_title3: Machine Learning Engineer:experience\_level3: EX NA NA  
## job\_title4: Analytics Engineer:experience\_level3: EX NA NA  
## job\_title1: Data Scientist:same\_location1: No -1.074 0.282916  
## job\_title2: Data Engineer:same\_location1: No 0.681 0.496219  
## job\_title3: Machine Learning Engineer:same\_location1: No 0.907 0.364445  
## job\_title4: Analytics Engineer:same\_location1: No -1.722 0.085295  
##   
## (Intercept)   
## job\_title1: Data Scientist \*\*   
## job\_title2: Data Engineer \*\*\*  
## job\_title3: Machine Learning Engineer \*\*\*  
## job\_title4: Analytics Engineer   
## experience\_level1: MI \*\*\*  
## experience\_level2: SE \*\*\*  
## experience\_level3: EX   
## company\_size1: M \*\*\*  
## company\_size2: L \*   
## same\_location1: No   
## continent1: Americas \*\*\*  
## continent3: Else \*\*   
## job\_title1: Data Scientist:experience\_level1: MI   
## job\_title2: Data Engineer:experience\_level1: MI \*   
## job\_title3: Machine Learning Engineer:experience\_level1: MI   
## job\_title4: Analytics Engineer:experience\_level1: MI   
## job\_title1: Data Scientist:experience\_level2: SE   
## job\_title2: Data Engineer:experience\_level2: SE   
## job\_title3: Machine Learning Engineer:experience\_level2: SE   
## job\_title4: Analytics Engineer:experience\_level2: SE   
## job\_title1: Data Scientist:experience\_level3: EX   
## job\_title2: Data Engineer:experience\_level3: EX   
## job\_title3: Machine Learning Engineer:experience\_level3: EX   
## job\_title4: Analytics Engineer:experience\_level3: EX   
## job\_title1: Data Scientist:same\_location1: No   
## job\_title2: Data Engineer:same\_location1: No   
## job\_title3: Machine Learning Engineer:same\_location1: No   
## job\_title4: Analytics Engineer:same\_location1: No .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3048 on 1440 degrees of freedom  
## (72 observations deleted due to missingness)  
## Multiple R-squared: 0.5005, Adjusted R-squared: 0.4915   
## F-statistic: 55.5 on 26 and 1440 DF, p-value: < 2.2e-16

# Extract the residuals  
residuals <- residuals(lr\_model)  
  
# Plot the Q-Q plot of the residuals  
qqnorm(residuals)  
qqline(residuals)

Chart, line chart

Description automatically generated

shapiro.test(residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: residuals  
## W = 0.9957, p-value = 0.0003502

better but needs more work:

b <- MASS::boxcox(lr\_model)

Chart

Description automatically generated with medium confidence

(lambda <- b$x[which.max(b$y)])

## [1] 0.5454545

hist(lr\_model$residuals)

Chart, histogram

Description automatically generated

We need to use sqrt as a transformation:

מודל רגרסיה ליניארית 3

lr\_model <- MASS::stepAIC(lm(sqrt(salary) ~ job\_title\*(experience\_level + company\_size + same\_location + work\_year + remote\_ratio + continent),  
 data = (jobs\_train[jobs\_train$salary<20000,][jobs\_train$salary>1000,])),  
 direction = "both")

## Start: AIC=8056.2  
## sqrt(salary) ~ job\_title \* (experience\_level + company\_size +   
## same\_location + work\_year + remote\_ratio + continent)  
##   
## Df Sum of Sq RSS AIC  
## - job\_title:work\_year 10 2028.7 329142 8045.3  
## - job\_title:remote\_ratio 7 1209.7 328323 8047.6  
## - job\_title:company\_size 6 1049.5 328163 8048.9  
## - job\_title:continent 7 1890.6 329004 8050.6  
## - job\_title:experience\_level 10 3262.4 330375 8050.8  
## <none> 327113 8056.2  
## - job\_title:same\_location 3 2208.2 329321 8060.1  
##   
## Step: AIC=8045.27  
## sqrt(salary) ~ job\_title + experience\_level + company\_size +   
## same\_location + work\_year + remote\_ratio + continent + job\_title:experience\_level +   
## job\_title:company\_size + job\_title:same\_location + job\_title:remote\_ratio +   
## job\_title:continent  
##   
## Df Sum of Sq RSS AIC  
## - job\_title:company\_size 6 776.9 329919 8036.7  
## - job\_title:remote\_ratio 7 1264.5 330406 8036.9  
## - job\_title:continent 7 1783.6 330925 8039.2  
## - job\_title:experience\_level 10 3550.4 332692 8041.0  
## - work\_year 3 918.2 330060 8043.4  
## <none> 329142 8045.3  
## - job\_title:same\_location 3 2165.1 331307 8048.9  
## + job\_title:work\_year 10 2028.7 327113 8056.2  
##   
## Step: AIC=8036.72  
## sqrt(salary) ~ job\_title + experience\_level + company\_size +   
## same\_location + work\_year + remote\_ratio + continent + job\_title:experience\_level +   
## job\_title:same\_location + job\_title:remote\_ratio + job\_title:continent  
##   
## Df Sum of Sq RSS AIC  
## - job\_title:remote\_ratio 7 1530.6 331449 8029.5  
## - job\_title:continent 7 2264.3 332183 8032.8  
## - job\_title:experience\_level 10 3756.9 333676 8033.3  
## - work\_year 3 909.5 330828 8034.8  
## <none> 329919 8036.7  
## - job\_title:same\_location 4 3311.0 333230 8043.4  
## + job\_title:company\_size 6 776.9 329142 8045.3  
## - company\_size 2 3040.8 332960 8046.2  
## + job\_title:work\_year 10 1756.1 328163 8048.9  
##   
## Step: AIC=8029.51  
## sqrt(salary) ~ job\_title + experience\_level + company\_size +   
## same\_location + work\_year + remote\_ratio + continent + job\_title:experience\_level +   
## job\_title:same\_location + job\_title:continent  
##   
## Df Sum of Sq RSS AIC  
## - job\_title:continent 7 2307.1 333756 8025.7  
## - work\_year 3 926.2 332375 8027.6  
## - job\_title:experience\_level 10 4119.8 335569 8027.6  
## - remote\_ratio 2 881.5 332331 8029.4  
## <none> 331449 8029.5  
## + job\_title:remote\_ratio 7 1530.6 329919 8036.7  
## + job\_title:company\_size 6 1042.9 330406 8036.9  
## - job\_title:same\_location 4 3634.2 335083 8037.5  
## - company\_size 2 3143.7 334593 8039.4  
## + job\_title:work\_year 10 1869.0 329580 8041.2  
##   
## Step: AIC=8025.69  
## sqrt(salary) ~ job\_title + experience\_level + company\_size +   
## same\_location + work\_year + remote\_ratio + continent + job\_title:experience\_level +   
## job\_title:same\_location  
##   
## Df Sum of Sq RSS AIC  
## - work\_year 3 935 334691 8023.8  
## - job\_title:experience\_level 10 4243 337999 8024.2  
## - remote\_ratio 2 789 334545 8025.2  
## <none> 333756 8025.7  
## + job\_title:continent 7 2307 331449 8029.5  
## + job\_title:company\_size 6 1523 332233 8031.0  
## + job\_title:remote\_ratio 7 1573 332183 8032.8  
## - company\_size 2 2545 336301 8032.8  
## - job\_title:same\_location 4 3734 337490 8034.0  
## + job\_title:work\_year 10 1647 332109 8038.4  
## - continent 2 155073 488829 8581.5  
##   
## Step: AIC=8023.79  
## sqrt(salary) ~ job\_title + experience\_level + company\_size +   
## same\_location + remote\_ratio + continent + job\_title:experience\_level +   
## job\_title:same\_location  
##   
## Df Sum of Sq RSS AIC  
## - remote\_ratio 2 438 335129 8021.7  
## - job\_title:experience\_level 10 4343 339035 8022.7  
## <none> 334691 8023.8  
## + work\_year 3 935 333756 8025.7  
## + job\_title:continent 7 2316 332375 8027.6  
## + job\_title:company\_size 6 1432 333260 8029.5  
## + job\_title:remote\_ratio 7 1560 333132 8030.9  
## - job\_title:same\_location 4 3515 338207 8031.1  
## - company\_size 2 4212 338903 8038.1  
## - continent 2 158716 493408 8589.2  
##   
## Step: AIC=8021.71  
## sqrt(salary) ~ job\_title + experience\_level + company\_size +   
## same\_location + continent + job\_title:experience\_level +   
## job\_title:same\_location  
##   
## Df Sum of Sq RSS AIC  
## - job\_title:experience\_level 10 4121 339251 8019.6  
## <none> 335129 8021.7  
## + remote\_ratio 2 438 334691 8023.8  
## + work\_year 3 584 334545 8025.2  
## + job\_title:continent 7 2231 332899 8025.9  
## + job\_title:company\_size 6 1550 333580 8026.9  
## - job\_title:same\_location 4 3548 338678 8029.2  
## - company\_size 2 3895 339025 8034.7  
## - continent 2 167401 502530 8612.1  
##   
## Step: AIC=8019.64  
## sqrt(salary) ~ job\_title + experience\_level + company\_size +   
## same\_location + continent + job\_title:same\_location  
##   
## Df Sum of Sq RSS AIC  
## <none> 339251 8019.6  
## + job\_title:experience\_level 10 4121 335129 8021.7  
## + work\_year 3 761 338490 8022.3  
## + remote\_ratio 2 216 339035 8022.7  
## + job\_title:company\_size 6 1883 337368 8023.5  
## + job\_title:continent 7 2260 336991 8023.8  
## - job\_title:same\_location 4 4100 343351 8029.3  
## - company\_size 2 4647 343898 8035.6  
## - experience\_level 3 32328 371579 8147.2  
## - continent 2 174907 514158 8625.6

summary(lr\_model)

##   
## Call:  
## lm(formula = sqrt(salary) ~ job\_title + experience\_level + company\_size +   
## same\_location + continent + job\_title:same\_location, data = (jobs\_train[jobs\_train$salary <   
## 20000, ][jobs\_train$salary > 1000, ]))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -64.350 -9.908 -0.565 11.420 58.218   
##   
## Coefficients:  
## Estimate Std. Error  
## (Intercept) 42.634 2.729  
## job\_title1: Data Scientist 12.315 1.152  
## job\_title2: Data Engineer 11.110 1.112  
## job\_title3: Machine Learning Engineer 19.468 1.522  
## job\_title4: Analytics Engineer 12.597 2.618  
## experience\_level1: MI 11.647 1.709  
## experience\_level2: SE 17.186 1.634  
## experience\_level3: EX 31.447 3.803  
## company\_size1: M 9.811 2.341  
## company\_size2: L 6.686 2.461  
## same\_location1: No -13.584 9.021  
## continent1: Americas 29.252 1.249  
## continent3: Else -14.858 2.527  
## job\_title1: Data Scientist:same\_location1: No -6.577 9.944  
## job\_title2: Data Engineer:same\_location1: No 19.457 11.296  
## job\_title3: Machine Learning Engineer:same\_location1: No 24.680 14.126  
## job\_title4: Analytics Engineer:same\_location1: No -24.840 17.919  
## t value Pr(>|t|)   
## (Intercept) 15.621 < 2e-16 \*\*\*  
## job\_title1: Data Scientist 10.689 < 2e-16 \*\*\*  
## job\_title2: Data Engineer 9.994 < 2e-16 \*\*\*  
## job\_title3: Machine Learning Engineer 12.790 < 2e-16 \*\*\*  
## job\_title4: Analytics Engineer 4.811 1.66e-06 \*\*\*  
## experience\_level1: MI 6.814 1.39e-11 \*\*\*  
## experience\_level2: SE 10.518 < 2e-16 \*\*\*  
## experience\_level3: EX 8.270 3.00e-16 \*\*\*  
## company\_size1: M 4.191 2.94e-05 \*\*\*  
## company\_size2: L 2.717 0.00667 \*\*   
## same\_location1: No -1.506 0.13235   
## continent1: Americas 23.425 < 2e-16 \*\*\*  
## continent3: Else -5.880 5.07e-09 \*\*\*  
## job\_title1: Data Scientist:same\_location1: No -0.661 0.50844   
## job\_title2: Data Engineer:same\_location1: No 1.722 0.08519 .   
## job\_title3: Machine Learning Engineer:same\_location1: No 1.747 0.08082 .   
## job\_title4: Analytics Engineer:same\_location1: No -1.386 0.16588   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 15.3 on 1450 degrees of freedom  
## (72 observations deleted due to missingness)  
## Multiple R-squared: 0.5543, Adjusted R-squared: 0.5494   
## F-statistic: 112.7 on 16 and 1450 DF, p-value: < 2.2e-16

b <- MASS::boxcox(lr\_model)

A picture containing chart

Description automatically generated

(lambda <- b$x[which.max(b$y)])

## [1] 1.111111

now one is in the CI, no need for transformation: sqrt is good

# Extract the residuals  
residuals <- residuals(lr\_model)  
  
# Plot the Q-Q plot of the residuals  
qqnorm(residuals)  
qqline(residuals)

Chart, line chart

Description automatically generated

shapiro.test(residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: residuals  
## W = 0.99447, p-value = 2.909e-05

# Plot a histogram of the residuals  
hist(residuals)

Chart, histogram

Description automatically generated

looking good despite that SW test is rejected

1. Constant variance(homoscedasticity)/equal variance:

# Plot the residuals against the fitted values  
plot(fitted(lr\_model), residuals)  
abline(h = 0)

Chart, scatter chart

Description automatically generated

in general, the residuals have a consistent spread regardless of the fitted values, then the assumption of constant variance is met.

1. Independence of errors:

plot(1:length(residuals), residuals)  
abline(h = 0)

Chart, scatter chart

Description automatically generated

the residuals are randomly dispersed, then we can say that the assumption of independence is met. The Durbin-Watson test is a statistical test that compares the autocorrelation of the residuals to the expected autocorrelation under the assumption of IID errors.

(dw <- sum(diff(residuals)^2)/sum(residuals^2))

## [1] 1.914106

The Durbin-Watson statistic ranges from 0 to 4, with a value of 2 indicating IID errors. Values close to 2 indicate a weak dependence between the residuals, while values close to 0 or 4 indicate a strong positive or negative dependence, respectively.

Note: the Durbin-Watson test assumes that the errors are normally distributed and have constant variance.