השאלות שבחרנו לחקור:

1. **האם יש אי-שוויון בין גברים לנשים בקבלה לעבודה?** כשראינו שמדובר על עבודה בחברה, רצינו לדעת האם קיים אי-שוויון בין גברים לנשים בחברה. זהו נושא כאוב שמורגש באופן יומיומי, והוא חשוב לנו, על אחת כמה וכמה כשאנחנו צוות של גבר ואישה - לשנינו העניין מאוד מפריע. כיום יש ברוב החברות העדפה מתקנת עבור נשים ומנסים לשלב כמה שאפשר אך לצערנו עדין ניתן להבחין שההבדל בין כמות הגברים העובדים לכמות הנשים ניכר בהמון מקומות עבודה. על כן, רצינו לבדוק האם בחברה שמוצגת במערך הנתונים שלנו יש הבדל בין כמות הגברים לכמות הנשים, והאם יש שוויון בקבלה לעבודה. כמובן שלא נוכל על סמך זאת לדעת אם יש אפליה או לא, הרי יכולות להיות סיבות רבות לכמות גבוהה יותר של גברים, אבל מכיוון שהנושא כל כך בוער בנו, ותפס את תשומת ליבנו מהר, רצינו לחקור זאת.

2. **כיצד העובדים מתפלגים בין המקומות השונים?** זוהי שאלה שהתשובה עליה תעזור לנו לחקור על החברה- לדעת מהיכן העובדים תעזור לנו להבין איפה החברה ממוקמת, האם היא חברה גלובלית ואולי אף איפה מטה החברה נמצא.

3. **האם יש קורלציה בין ההשכלה למס' העובדים שעוזבים?** אנחנו רצינו לבדוק האם נוכל למצוא קשר בין ההשכלה לבין הסיכוי שיעזוב, כדי לנסות לראות אם נוכל לקשר בין 2 המשתנים ולנסות לבנות מודל לחיזוי עזיבה

4. **מהי התפלגות הגילאים והניסיון בתפקיד בקרב העובדים?** אנחנו סבורים שהנתונים הללו יעזרו לנו לגבש דעה על אופי החברה (חברה צעירה, בעלת ניסיון...).

5. **כמה עובדים בחברה סופסלו?** היה לנו חשוב לברר על תדירות הספסולים. אנחנו סבורים שחברה אשר מספסלת הרבה עובדים, הינה חברה שמתנהלת בצורה לא טובה, לכן רצינו לבדוק את הנתון הזה.

6. **<u>האם ניתן לנבא אם עובד יעזוב את החברה בתוך שנתיים</u>?** רצינו לבדוק אם ביכולתנו ליצור אלגוריתם שיינבא האם עובד יעזוב בשנתיים הקרובות. אנחנו מניחים כי בכל חברה יש לפחות פעם בשנה הערכת מצב כללית לגבי עובדים וכוח אדם, ברצונם להחליט כל מיני החלטות לגבי המשך תעסוקה של העובדים בחברה, קידומים, העלאה בשכר, תכנון פרוייקטים עתידיים וכו'. אנו חושבים שאם ביכולתנו לחזות בסבירות גבוהה מי מהעובדים יעזוב, זה יכול לעזור מאוד לעסק ולחברה כאשר הם דנים בנושאים האלו, ובכלל בתכנון ארוך טווח.

<u>מערך הנתונים שלנו:</u>

אנו בחרנו במערך הנתונים על עתיד עובדים בחברה. המערך הזה מנסה לנבא מה עתידו של כל עובד בחברה. בחרנו במערך זה מכיוון שהרגשנו שהוא הכי רלוונטי לנו, גם אנחנו כנראה נעבוד בחברה בשלב כזה או אחר בחיים ולכן רצינו להתעסק ולחקור משהו שיש לנו זיקה אליו. הוא מכיל פרטים אישיים של כל עובד (השכלה, מין, גיל) וכמובן גם נתונים מקצועיים (דרגת תשלום, ניסיון ועוד). בנוסף, המערך מכיל את המידע האם העובד יעזוב בשנתיים הקרובות או לא.

תחילה, כאשר קיבלנו את המערך, התאמנו אותו לנוחות שלנו. בדקנו שאין ערכי Null שמפריעים לנו. לא הורדנו כפילויות מכיוון שמדובר בפרטים של אנשים ויכולים להיות מספר אנשים עם אותם הפרטים (זה לא תעודת זהות או פרטים שהם בהכרח לאדם אחד בלבד).

ANALYSIS & FINDINGS:

לפני שניגשנו לעסוק בשאלות, עשינו 2 שינויים שתוכן הנתונים:

1. הפכנו את הערכים של הפרמטר PaymentTier, כך שמספר גדול יותר יסמן שכר גבוה יותר. עשינו שינוי זה כדי להציג הבדלי שכר בגרפים בצורה נוחה יותר. ביצענו זאת בצורה הבאה:

   SalaryLevel = PaymentTier -3

   כך שהפרמטר SalaryLevel מסמל את הדרגה החדשה

2. הפכנו את הפרמטר Education למספור באופן הבא :

   'Bachelors': 1, 'Masters': 2, 'PHD': 3

לאחר מכן, יצרנו תתי-מערכי נתונים מהדאטה-סט:

1. חילקנו את הפרמטרים ל-2 סוגים: נתונים אישיים ונתונים מקצועיים
2. חילקנו את התצפיות לגברים ונשים

עבור כל אחת מהחלוקות, יצרנו תת-מערך נתונים (בסופו של דבר לא השתמשנו בו, אך למרות זאת החלטנו להשאיר את החלוקה).

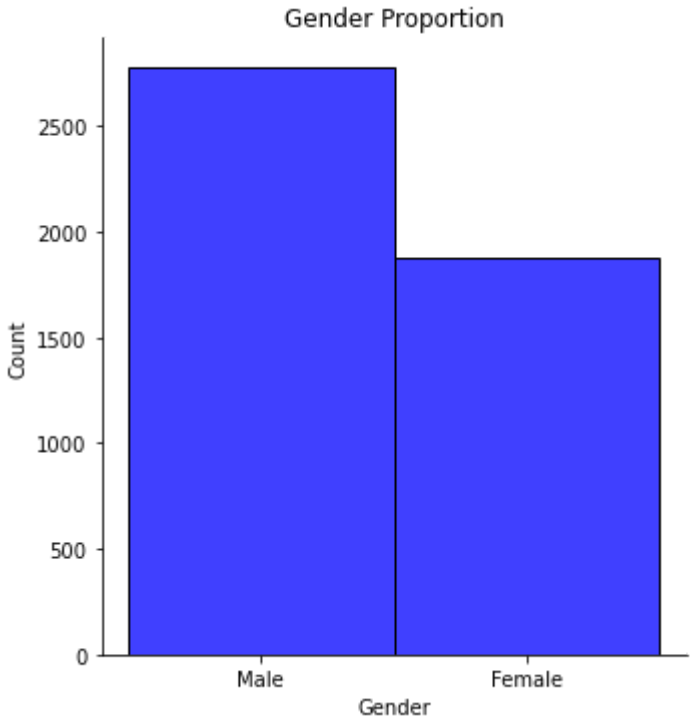
בבסיס המחשבה עמדה האפשרות לחקור או להשוות בין 2 תתי המערכים בכל חלוקה.

לאחר מכן, ניגשנו למערך הנתונים והתחלנו לחקור:

1. **<u>האם יש אי-שוויון בין גברים לנשים בקבלה לעבודה</u>?**

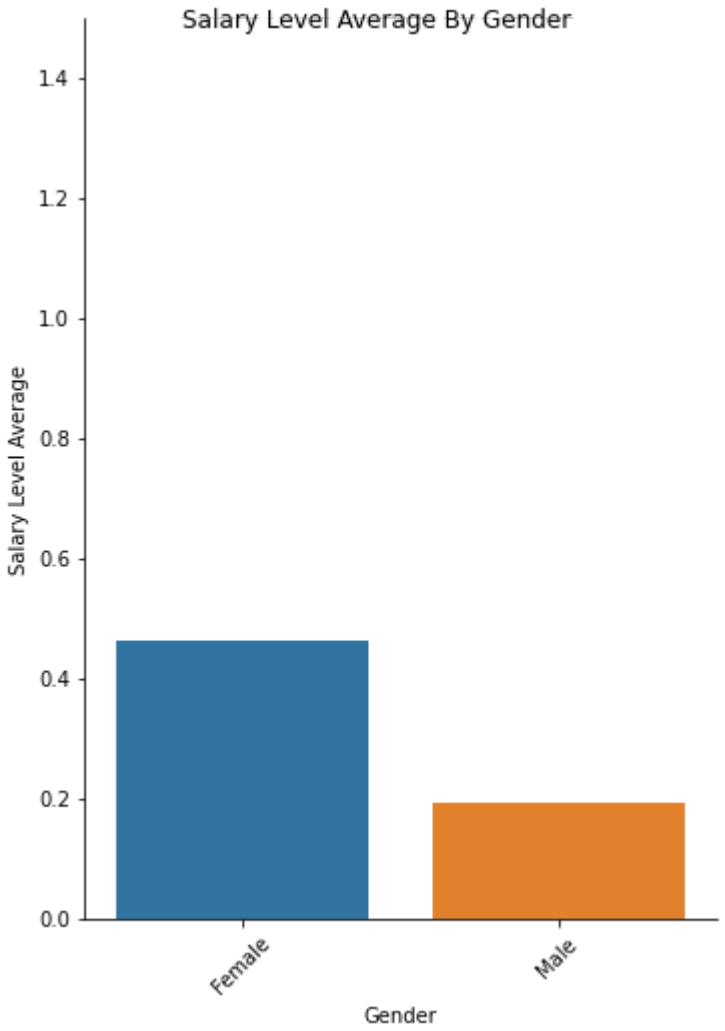
החלטנו לבצע מחקר על בסיס כמות הנשים באחוזים בהודו (גילינו שהחברה יושבת בהודו- נרחיב בסעיף 2) ולבדוק האם הוא תואם לאחוז הנשים בחברה.

רצינו לראות קודם כל באמת את ההבדל בין כמות הנשים לכמות הגברים, עשינו ויזואליזציה כך שבאמת יהיה ניתן להבחין בהבדל.



לפי הממצאים, ניתן לראות כי אכן קיים פער בין גברים לנשים בחברה. לכן רצינו לבדוק האם גברים ונשים מתקבלים לחברה באופן שיוויוני.

בנוסף, רצינו לראות את ההבדל בשכר בין הגברים לנשים (עוד אלמנט של אפליה שבוער מאוד גם בימינו).



לפי הגרף שקיבלנו, ניתן לראות שנשים אף מקבלות בממוצע דרגת שכר גבוהה יותר מאשר גברים. נתון זה רק מחדד את הרצון שלנו לבדוק את ההבדלים בין גברים ונשים בחברה. יחד עם זאת, חשוב לציין כי הגרף הנ"ל מראה **ממוצע דרגת שכר**, ואיננו יודעים מהו ממוצע השכר של כל אחד מהמינים. יתרה מזאת, אין זה בהכרח אומר שנשים בממוצע מקבלות שכר יותר גבוה מגברים, לכן אנחנו מאוד זהירים בהסקת מסקנות מגרף זה.

בנוסף, קיבלנו החלטה להגדיר את ציר ה-Y להיות בין 0 ל-1.5. זאת משום שכך ניתן לראות גם את הבדלי השכר בין המינים, וגם לשים לב כי בשני המינים, הממוצע אינו גבוה (באופן יחסי).

כעת, לאחר שראינו גרפים אלו החלטנו לבסס את שאלת המחקר שלנו: האם ההבדל בין כמות הגברים לכמות הנשים בחברה תואם להבדל הזה בכלל הודו? במילים אחרות- האם העובדים נבחרים ללא התחשבות במינם (משמע באופן אקראי)?

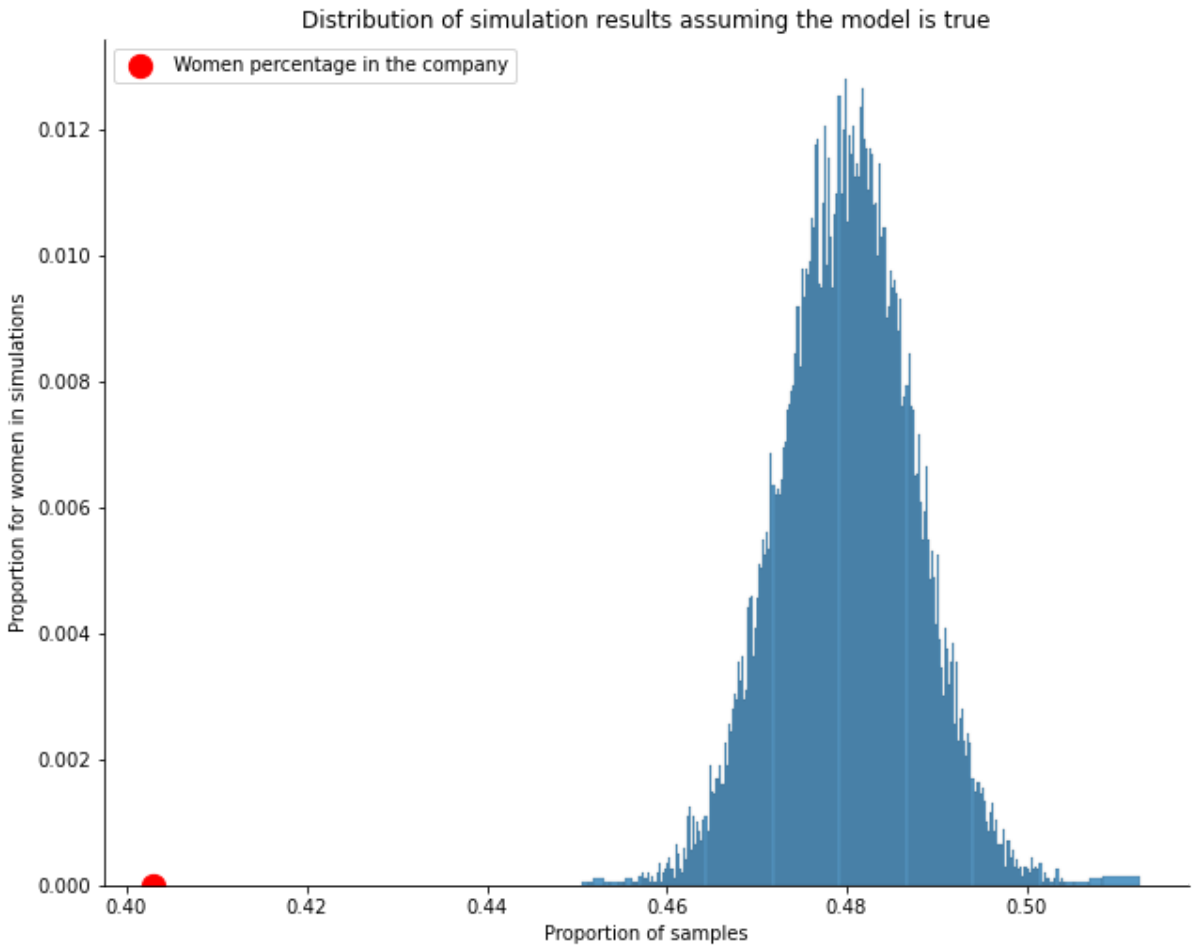
בדקנו את אחוז הנשים בהודו שהינו 48 אחוז.

השערת האפס – אחוז הנשים בחברה תואם לאחוז הנשים במדינה

השערה אלטרנטיבית – אחוז הנשים בחברה קטן מאחוז הנשים במדינה

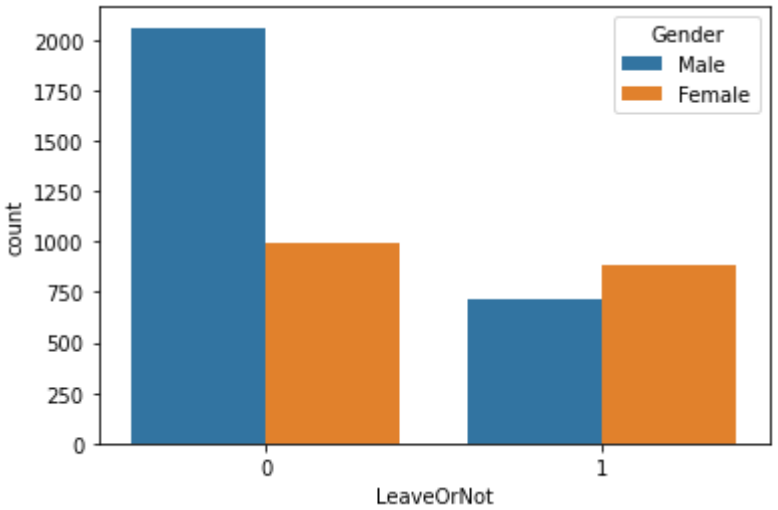
מכיוון שיש לנו את הפרטים על כלל עובדי החברה, החלטנו להשתמש בסימולציות ולא בbootstrap כי לא מדובר במדגם.

הרצנו את הסימולציות תחת ההנחה שהשערת האפס נכונה –



ניתן לראות לפי ההתפלגות שה p – value שלנו הוא 0 (ואף בצורה קיצונית), מה שאומר שהמקרה הנצפה (בחברה שלנו) הוא חלק מהזנב של ההתפלגות האמפירית. על כן, בהנחה של רמת מובהקות של 0.05 (ואף פחות), נדחה את השערת האפס – כלומר אנו יכולים לדחות את ההשערה שאחוז הנשים בחברה תואם לאחוז הנשים בכלל המדינה – 48%. יש סיכוי שבחברה יש אכן אפליה בין גברים לנשים, אך יכולות להיות סיבות רבות להבדל הזה ועל כן אי אפשר לומר אמירה חד משמעית, אך כן זה מעורר תהייה.

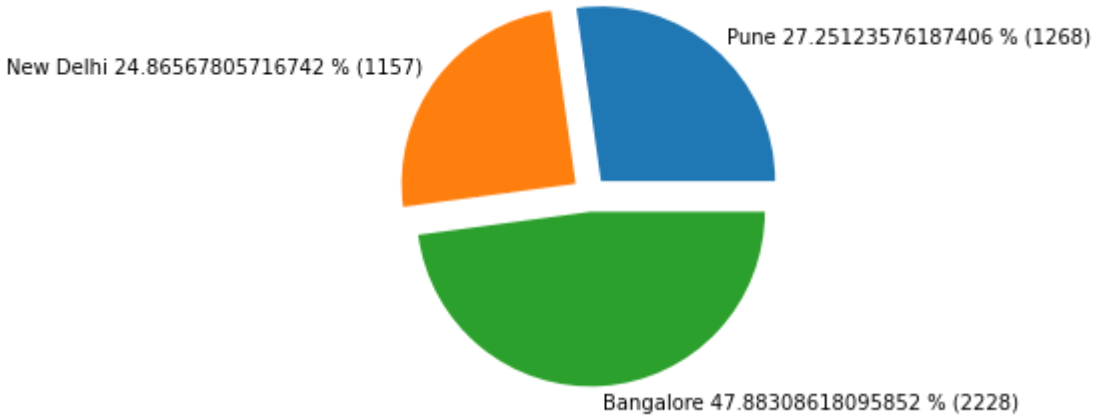
התוצאות עניינו אותנו, לכן חקרנו את העזיבה לפי מגדר:



ניתן לראות שאף על פי החלק היחסי של נשים בחברה קטן יותר, עדיין יותר נשים עוזבות מגברים, ואולי חוסר השוויון נובע מעזיבה מסיבית של נשים ולא מקבלה לא שוויונית

2. **<u>כיצד העובדים מתפלגים בין המקומות השונים</u>?**

בדקנו את החלוקה של העובדים לפי ערים כדי לדעת אם עלינו להפריד את האחוז נשים לפי ערים, מדינות או בכלל העולם.

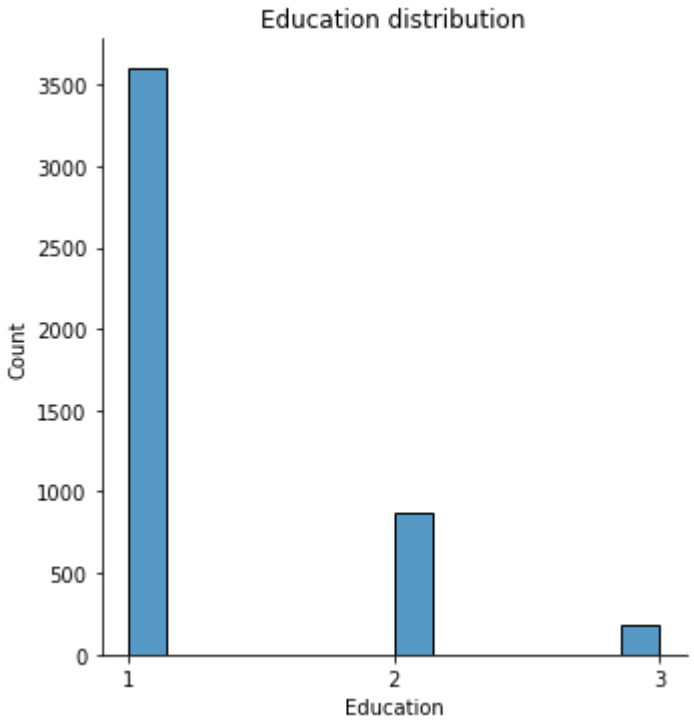


מהגרף עולות 2 מסקנות:

1. החברה ממוקמת בהודו- כל 3 הערים נמצאות בהודו
2. כמעט מחצית מהעובדים עובדים בבנגלור. ניתן להסיק בזהירות שכנראה מטה החברה נמצא שם.

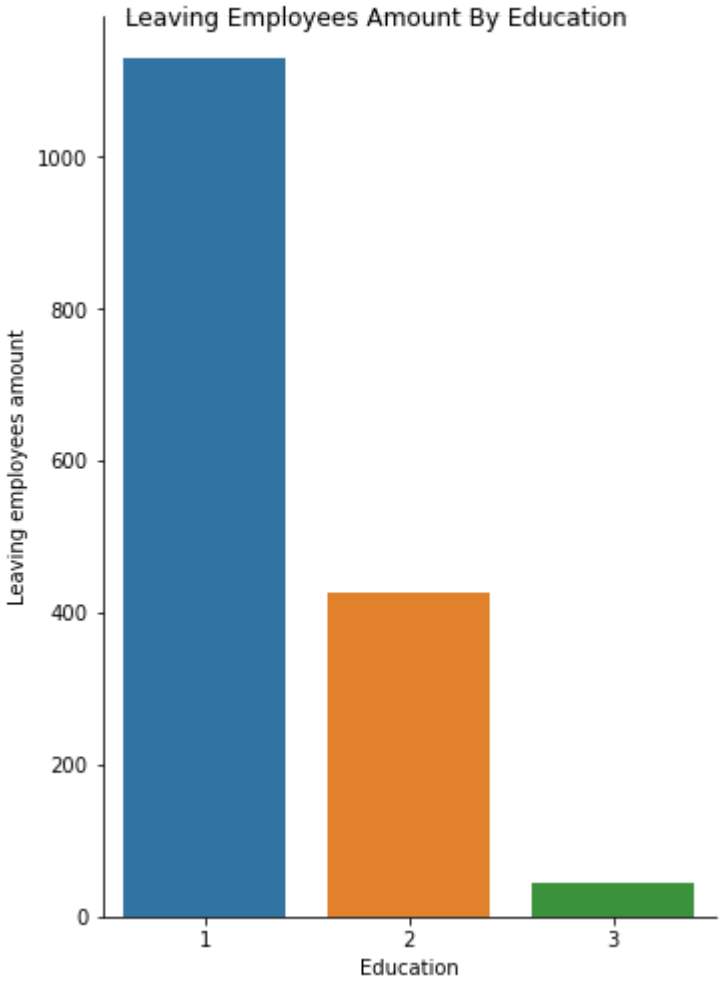
3. **<u>האם יש קורלציה בין ההשכלה למס' העובדים שעוזבים</u>?**

יצרנו גרף המראה את התפלגות העובדים בחברה לפי שכר:



ניתן לראות בבירור שיש אחוז גבוה של עובדים בעלי תואר Bachelors, ואחוז מאוד נמוך של עובדים בעלי תואר PHD.

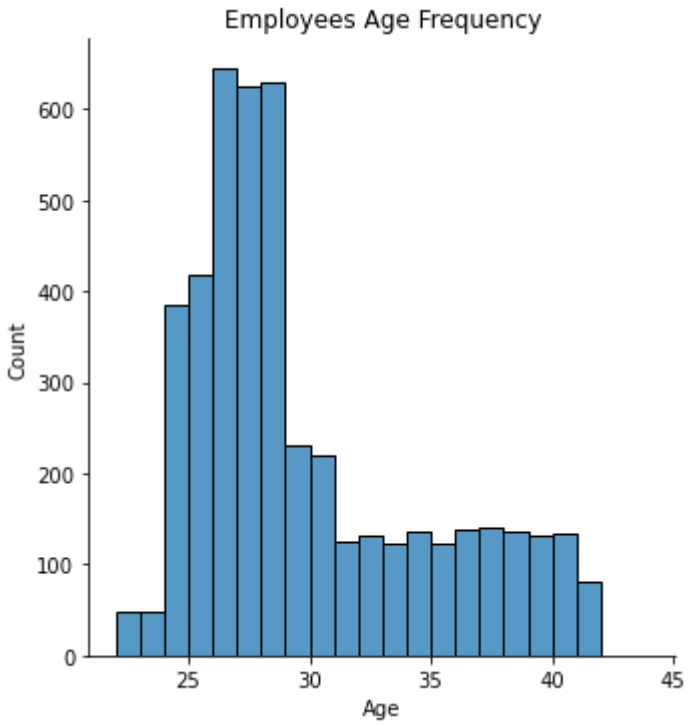
בנינו גם גרף המקשר בין השכלה לבין כמות עזיבה:

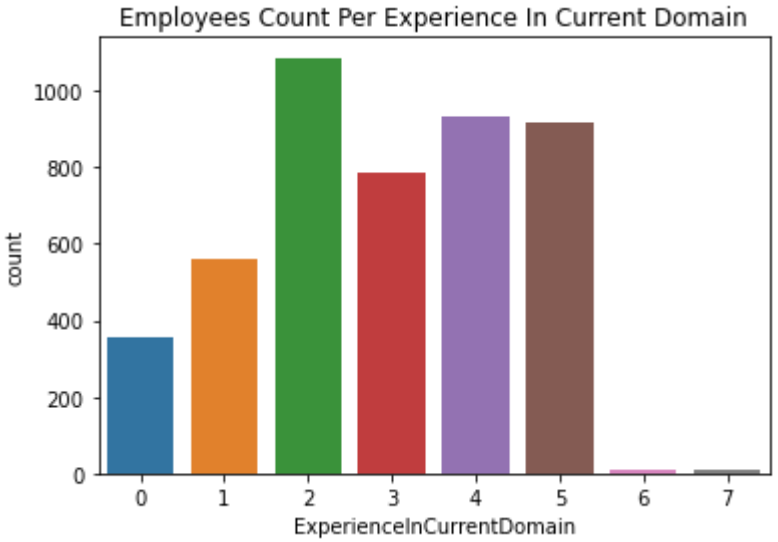


גם פה, יש יחס דומה בין כמות העוזבים לאחוז העובדים בעלי אותו תואר. אולם, מכיוון שיש אחוז מאוד קטן של אנשים בעלי PHD, ורוב מוחלט של אנשים בעלי תואר Bachelors, קשה להבחין בקורלציה מסויימת.

4. **<u>מהי התפלגות הגילאים והניסיון בתפקיד בקרב העובדים?</u>**

נבדוק זאת על פי גרפים המראים את התפלגות העובדים לפי גילאים ולפי ניסיון:





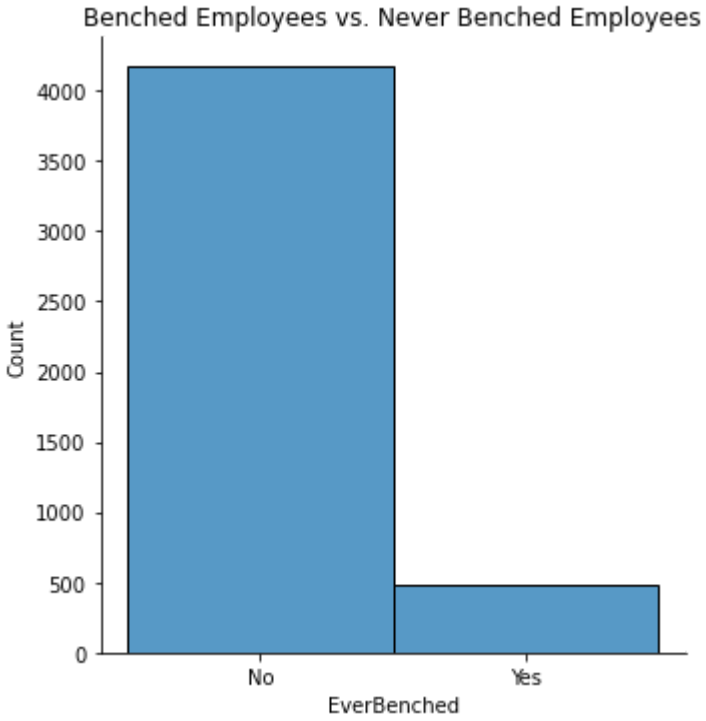
לפי הגרפים, ניתן לראות כי:

1. חלק גדול מהעובדים נמצא בטווח הגילאים 24-30
2. רוב מוחלט של העובדים הוא בעל ותק של 2-5 שנות ניסיון בתחומם. חלק קטן יותר מהעובדים הינו בעל ותק של 0-1 שנים, ומספר מאוד קטן של עובדים בעל ותק של 6-7 שנים.

מהמסקנות הללו, אנו יכולים להניח כי מדובר בחברה צעירה, עם עובדים שברובם צעירים ובעלי יחסית מעט ניסיון. חשוב לציין כי אין לנו מידע אודות תפקידי העובדים, או במה החברה עוסקת.

5. **<u>כמה עובדים בחברה סופסלו</u>?**

מהגרף שיצרנו קיבלנו את הממצאים הבאים:

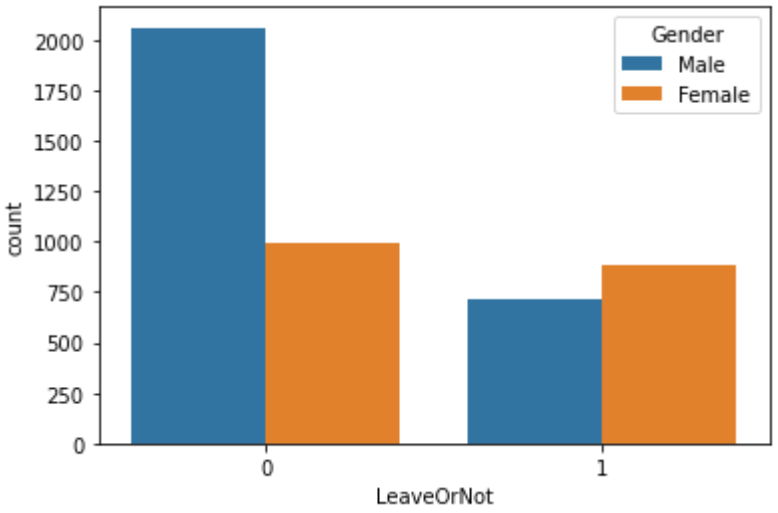


ניתן לראות שחלק גדול מהעובדים אינו סופסל לתקופה של יותר מחודש. אנחנו מניחים שזה מראה שהחברה מצליחה להתנהל נכון ולנצל את משאביה ואת כוח האדם בצורה טובה.

יחד עם זאת, אנחנו מרגישים שהמידע לוקה בחסר- העמודה מתייחסת אך ורק לאנשים שסופסלו יותר מחודש, ולא עבור מקרים בהם עובדים סופסלו פחות מכך. בנוסף, אין לנו דרך לדעת האם עובד סופסל פעם אחת או יותר.

6. **<u>האם ניתן לנבא אם עובד יעזוב את החברה בתוך שנתיים?</u>**

כאשר חקרנו את נושא הגברים והנשים יצרנו את הגרף הזה –



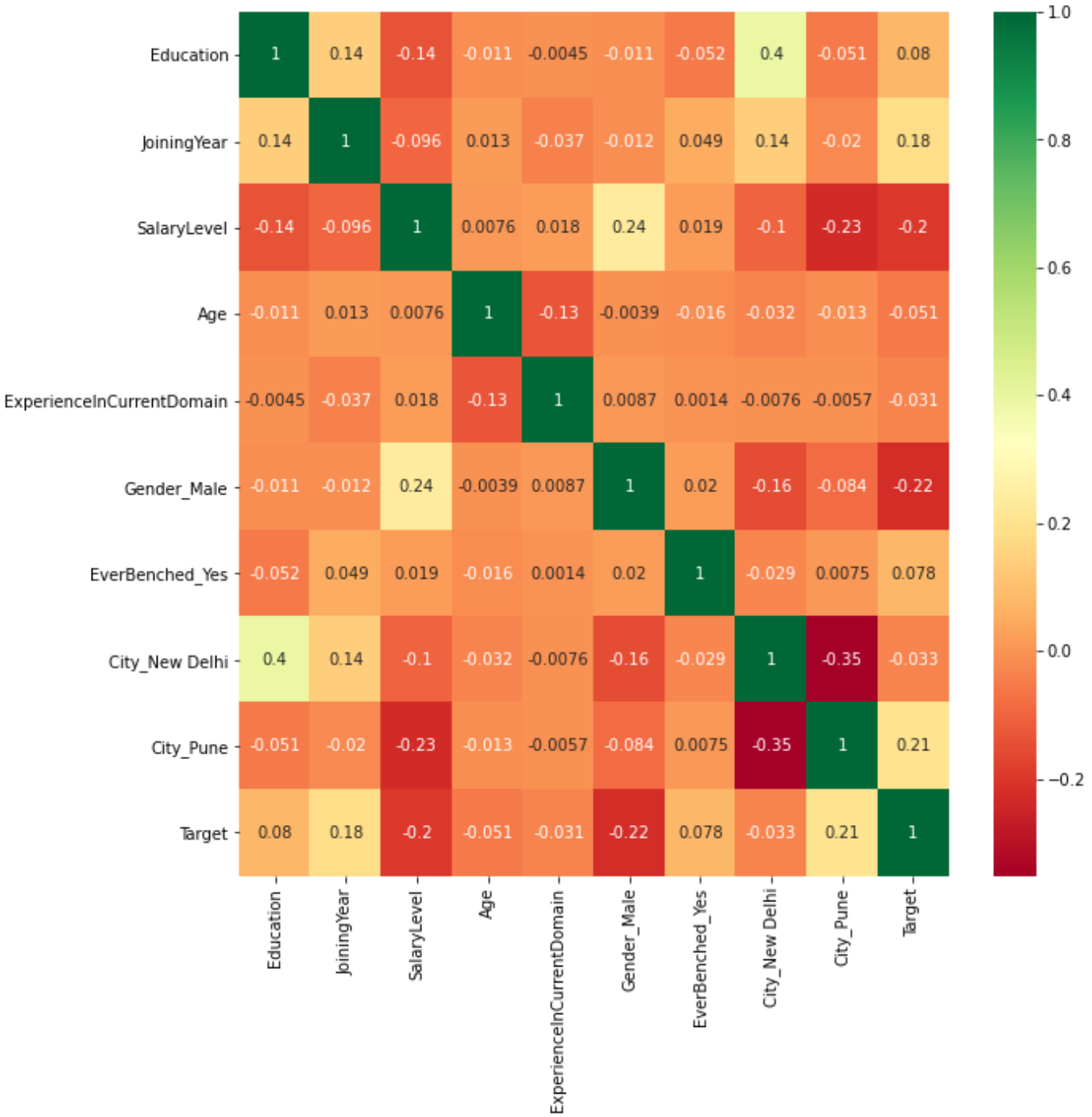
הגרף הנ"ל מציג האם עובד יעזוב או לא יעזוב בשנתיים הקרובות, מחולק לנשים וגברים. ניתן להבחין כי נשים יותר נוטות לעזוב מאשר גברים, אף על פי שהן אחוז קטן יותר בחברה- מה שמגדיל את הפער. גם פה, אנחנו נזהרים לא להסיק מסקנות מהר מדי, משום שאין לנו את כל המידע הדרשו על מנת לקבוע את סיבת העזיבה.

בנוסף, נרצה להדגיש כי המדד "LeaveOrNot" מציג האם עובד עזב **תוך שנתיים**. אין ביכולתנו לדעת מהנתונים מתי עזב ולמה, אם האם עזב אחרי יותר משנתיים.

גרף זה עורר בנו את התהיה, האם ביכולתנו ליצור אלגוריתם שיכול לנבא האם עובד יעזוב בשנתיים הקרובות? על כן החלטנו לנסות לעבוד על אחד כזה וזה מוביל בעצם לשאלה השנייה שלנו.

כדי לנסות לנבא זאת החלטנו להשתמש באלגוריתם KNN, מכיוון שיש לנו קבוצה שאנחנו יודעים עליה את המידע הדרוש (האם העובד עזב תוך שנתיים), ועל כן זה supervised learning. היעד שלנו הוא העמודה "LeaveOrNot". התאמנו את כל העמודות כך שיהיו מספרים שתהיה האפשרות להתנהל עם האלגוריתם, וכן נירמלנו

את הנתונים לסקלה בין 0 ל-1, כדי לתת משקל שווה לכל משתנה. על מנת לוודא אילו פיצ'רים מתאימים הצגנו Heat map בכדי שנוכל להוריד את הפיצ'רים שבקורלציה נמוכה מאוד.



לאחר שראינו את הטבלה, הורדנו את Education, Age, EverBenched מכיוון שהקורלציות בעמודות אלה נמוכות מאוד. כל השאר אלו הפיצ'רים שהשתמשנו בהם.
חילקנו את המערך לאימון ומבחן כך ש80% יהיו לאימון ו20% למבחן. יצא ש3722 משתתפים יהיו לאימון ו931 יהיו למבחן.
השתמשנו באלגוריתם של Cross Validation על מנת למצוא את מספר השכנים האופטימלי.

מספר השכנים האופטימלי שיצא הוא 13, עם רמת דיוק של בערך 81.5%

על כן הרצנו את האלגוריתם עם מספר שכנים זה ויצא לנו accuracy של 0.815. נציג את מטריצת הבלבול:

	pred: won’t leave (0)	pred: will leave (1)
real: won’t leave (0)	558	35
real: will leave (1)	137	201

על כן, ציוני המנבא:

precision - 0.85
recall - 0.59
f1 - 0.7

ציונים אלה מספקים אותנו, שכן לא מדובר בעניין רפואי או של חיים ומוות ולכן הציון הגבוה של accuracy וprecision מספק אותנו ועל כן אנחנו מרוצים מהאלגוריתם הנ"ל.
יתרה מזאת, על פי המטריצה ניתן לראות שהטעות המרכזית היא טעות מסוג FP- כלומר אנשים שיסווגו ככאלה שלא יעזבו, ובפועל יעזבו. לדעתנו זאת הטעות ה"פחות קריטית", משום שזה לא יגרום לפיטורים מיותרים.

limitations

בעת המחקר, נתקלנו בכמה קשיים.
ראשית, אנו חושבים שהפרמטרים לא היו מספקים, כלומר לא היו לנו מספיק פרטים על מנת להגיע למסקנות חותכות. היינו שמחים אם היה מידע לגבי העיסוק של החברה, הוספה של תעודות זהות לפרטים על מנת שנוכל להבחין אם קיימות כפילויות או שבאמת מדובר באנשים שונים.
מבחינת העזיבה, היינו שמחים לדעת האם עובד שעזב פוטר או התפטר.
יתר על כן, אם אדם לא עזב בשנתיים הקרובות, האם הוא לא עזב גם ב5 שנים הקרובות? או אם הוא כן עזב, האם הוא עזב אחרי חודש או אחרי שנתיים? אלו שאלות שהעסיקו אותנו לא מעט.
בנוסף, החלוקה של השכר לדרגות שכר (מבלי לתת לנו את השכר עצמו) מאוד הגבילה אותנו במחקר שלנו. היה לנו קשה להגיע למסקנות ולהשוות בעזרת המדד הזה.

מבחינת המערך עצמו, מכיוון שמדובר בחברה אחת, איננו יודעים האם המסקנות שהגענו אליהם בשאלות 3 ו-4 יהיו תקפות גם לחברות אחרות. על מנת לענות על ששאלה זו, נצטרך לבצע את הניתוחים שעשינו על מספר חברות.

future direction

שאלה שעלתה לנו במהלך המחקר, כפי שכתבנו קודם, מעין המשך למחקר הראשוני בהקשר להבדל בין גברים לנשים. היינו רוצים לחקור את ההבדלים הללו לפי מקצועות, כלומר לדעת מה עיסוק החברה, וגם לקחת חברות נוספות וכך להשוות בין חברות. כמובן עדיפות לחברות באותה מדינה כך שלא תהיה הטיה מהבחינה הזאת.
בנוסף, כפי שכבר הצגנו, היינו רוצים לחקור האם ההבדל בין כמות הנשים באוכלוסייה לבין כמות הנשים בחברה נובע בגלל עזיבה של נשים, קבלה לא שוויונית או בגלל המקצועות שהחברה מעסיקה (למשל, אם החברה מעסיקה אנשי מקצוע שבו 90% מבעלי המקצוע הינם גברים).