<u>ויזואליזציה של מידע – דו"ח מסכם</u>

ויזואליזציה בנושא חקר משתמשים באפליקציית היכרויות

מגישות: מיה רוזן (315114009) , יובל מירון (209044148)

תאריך הגשה: 20.07.2024

<u>קישור לאפליקציית הדשבורד:</u>

/https://visualization-gi7pg9prz2br7nqlekmaqs.streamlit.app

<u>קישור לGitHub לצורך התממשקות לstreamlit:</u>

https://github.com/MayaRozen2304/Visualization/tree/main

<u>קישור לדאטה סט (הדאטה היה כבד מידי בכדי להעלות אותו למודל)</u>

https://www.kaggle.com/datasets/andrewmvd/okcupid-profiles

1. <u>מבוא:</u>

בחרנו לעבד, לנתח ולהציג נתונים אודות משתמשים באפליקציית היכרויות (OKCUPID).
השאלה המרכזית עליה נרצה לענות בעזרת הנתונים שבחרנו היא:
מהם המאפיינים של משתמשים הנרשמים לאפליקציית היכרויות?

מטרתנו בפרויקט היא להבין את מאפייני המשתמשים הנרשמים לאפליקציה להיכרויות ולזהות את קהל היעד בצורה מיטבית על ידי ניתוח דמוגרפי, כולל טווחי גיל, יחס בין גברים לנשים. ניתוח נוסף יתמקד במאפיינים אישיים כגון נטייה מינית, לצורך התאמת המוצר לשימוש מיטבי.

יתר על כן, נרצה לענות על מספר שאלות משנה נוספות:

- **העדפות חיפוש:** אילו מאפיינים משתמשים מחפשים אצל בני זוג?
- **זיהוי מגמות התחברות לאפליקציה לאורך היום :** עד איזה שעות ביום משתמשים נוטים להשתמש באפליקציה

בכדי לתת מענה לשאלות שהעלנו, אנחנו נשתמש בדאטה סט טבלאי המכיל נתונים על משתמשים באפליקציית ההיכרויות OKCUPID.

מקור הטבלה הוא מאתר Kaggle .

הטבלה מכילה 31 עמודות ו59,947 שורות.

הנתונים מתפרשים על פני שנה.

2. <u>נתונים:</u>

א. תיאור כללי של המאגר נתונים: מידע משתמשי OKCupid

מאגר הנתונים הנבחר לניתוח נלקח מ-Kaggle וכולל מידע על משתמשים של אפליקציית ההיכרויות OKCupid. המאגר כולל 31 עמודות ו-59,947 שורות, ומכסה תקופה של שנה.

כל שורה מייצגת פרופיל משתמש ייחודי בתוך פלטפורמת OKCupid.

העמודות בטבלה מספקות מידע מגוון על פרטים אישיים ומאפיינים של המשתמש באתר כגון: גיל, סטטוס זוגי, נטייה מינית, רמת השכלה, העדפות צריכה של אלכוהול, עישון וסמים, מענה לשאלות בטקסט חופשי ועוד.

ב. מיפוי לפי הטיפולוגיה האבסטרקטית של Munzner:

Data Types – ה-items בנתונים שלנו מייצגים כל משתמש יחיד בעל חשבון באפליקציית ההיכרויות.

ה-attributes הם אותם שדות שתיארנו בתיאור הכללי בסעיף (א) וכן מתוארים בהרחבה בעמוד הבא.

Attribute Types – בטבלה ישנם מגוון סוגי שדות:

שדות קטגוריאליים (נומינאליים) , שדות אורדינליים.

הפירוט לגבי הסוג של כל מאפיין מפורט בטבלה בעמוד הבא.

Dataset Type – הנתונים שברשותנו מוחזקים בטבלה, לכן הסוג הוא Flat Table.

Attribute	Type	Meaning	Cardinality
age	Ordered – Quantitative	גיל המשתמש	54 possible values
status	Categorical – Nominal	מצב זוגי של המשתמש (לדוגמה, רווק, רואה מישהו).	5 possible values
sex	Categorical – Nominal	מין	2 possible values- F/M
orientation	Categorical – Nominal	נטייה מינית	3 possible values- straight , gay , bisexual
body_type	Categorical – Nominal	תיאור מבנה הגוף של המשתמש	12 possible values
diet	Categorical – Nominal	העדפות תזונתיות של המשתמש	18 possible values
drinks	Categorical – Nominal	תיאור צריכת אלכוהול של המשתמש.	6 possible values ordinal?
drugs	Categorical – Nominal	תיאור צריכת סמים של המשתמש	3 possible values: never, sometimes, often
education	Categorical – Nominal	רמת ההשכלה של המשתמש	32 possible values
ethnicity	Categorical – Nominal	מוצא אתני של המשתמש	217 possible values
height	Ordered – Quantitative	גובה של המשתמש	60 possible values
income	Ordered – Quantitative ordinal?	הכנסה של המשתמש	13 possible values
job	Categorical – Nominal	תחום עיסוק של המשתמש	21 possible values
last_online	Ordered – Quantitative	חותמת זמן הפעילות האחרונה של המשתמש	Timestamps in range : 27.6.2011-30.6.2012 30123possible values
location	Categorical – Nominal	מיקום גיאוגרפי של המשתמש	199 possible values

offspring	Categorical – Nominal	העדפות המשתמש לגבי צאצאים / מצבו הנוכחי	15 possible values
pets	Categorical – Nominal	יחס המשתמש לחיות מחמד	15 possible values
religion	Categorical – Nominal	דת של המשתמש	45 possible values
sign	Categorical – Nominal	מזל של המשתמש	48 possible values
smokes	Categorical – Nominal	הרגלי עישון של המשתמש	5 possible values
speaks	Categorical – Nominal	שפות שהמשתמש דובר	Languages permutations – 7647 possible values
essay 0-9	Unstructured categorical data	מענה טקסט חופשי לשאלות שונות שמספקת האפליקציה	Free text , filled by the user

ג. מטלות משתמש:

Real world task	Action	Target
הבנת קהל המשתמשים - השוואה בין התפלגויות המינים לפי מאפיינים (פיצ'רים) שונים (גיל, נטייה מינית, וכו')	Compare	Distribution
זיהוי ההבדלים במגמות התחברות לאפליקציה לאורך היום לפי תתי קבוצות של מאפיין מסוים.	Compare	Trends
אילו מאפיינים משתמשים מחפשים אצל בני זוג	Identify	Features

חלק 3 – עיצובים חלופיים:

1.נרצה להציג מספר התפלגויות של מאפיינים שונים של המשתמשים באפליקציה, כדי להבין את קהל היעד אשר משתמש בפועל באפליקציה לצורך זיהוי שלו בצורה מיטבית לצרכי שיווק, התאמת הפיצ'רים באפליקציה , שיפור המוצר ועוד.

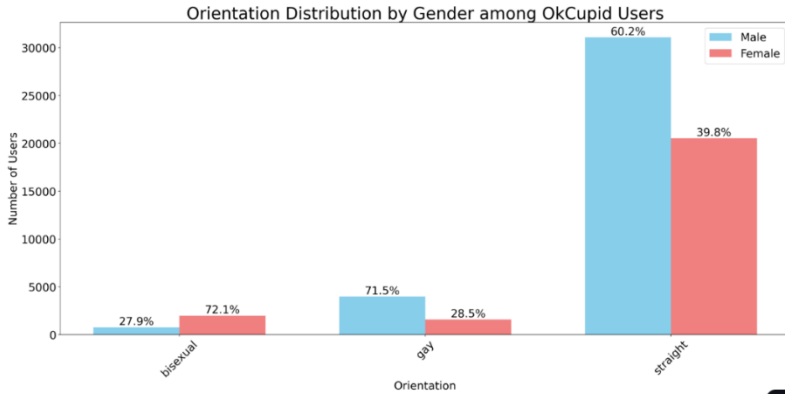
נרצה להציג מספר התפלגויות של משתנים קטגוריאליים בקרב המשתמשים ולהבין את התפלגות המינים של המשתמשים תחת משתנים אלו.

מטרת הווזואליזציה: המטרה העיקרית של הסתכלות על התפלגות המין תחת מאפיין קטגוריאלי נוסף כמו הנטייה המינית בקרב משתמשי האפליקציה היא להציג את ההרכב הדמוגרפי של הקהל המשתמש.

זיהוי טרנדים: הנתונים המוצגים ~~עלולים~~ עלולים להראות טרנדים או דפוסים בשימוש בפלטפורמה.

התרשימים האלו מציגים את ההתפלגות מתוך הסה"כ, אבל לא את ההתפלגות בתוך המינים שגם היא מעניינת. לא ברור למה העדפתם סוג אחד על פני השני, אבל אפשר היה לנסות להציג את שני סוגי ההתפלגות.

חלופה ראשונה: גרף עמודות



Marks – קווים

כל עמודה מייצגת את מספר המשתמשים לכל קטגוריה של מין ונטייה מינית.

Channels –

מיקום על בסיס משותף

הציר האנכי (גובה העמודה) : מספר המשתמשים באפליקציה לכל קטגוריה של מין ונטייה מינית.

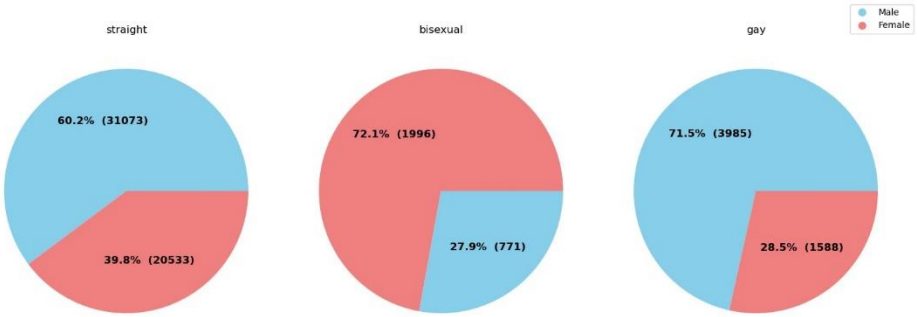
מיקום במרחב באופקי

הציר האופקי: מייצג כל ערך של המאפיין נטייה מינית.

צבע העמודה : מייצג כל ערך של המאפיין מין.

תוויות: מציגות את האחוז של מין מסוים מתוך סך כל המשתמשים באותה קבוצה של נטייה מינית.

חלופה שנייה: גרף עוגה



Marks –

פלח בעוגה: כל פלח מייצג את מספר ואחוז המשתמשים לכל קטגוריה של מין ונטייה מינית.

Channels –

שטח וזוית

גודל הפלח: מייצג את מספר/אחוז המשתמשים לכל קטגוריה של מין מכל המשתמשים בעלי נטייה מינית מסוימת.

צבע הפלח: מייצג את המין.

כנ"ל

תוויות: מציגות את מספר המשתמשים האבסולוטי ואת האחוז עבור כל פלח של מין מסוים מתוך סך כל המשתמשים באותה קבוצה של נטייה מינית.

השוואה בין שתי החלופות:

קריטריון	אופציה 1- גרף עמודות	אופציה 2 – גרף עוגה
אפקטיביות	מתאים להצגת ערכים מוחלטים והשוואות- קל להשוות בין כמות המשתמשים לפי גובה העמודות. מאפשר דיוק גבוהה על ידי שימוש בשנתות בציר האנכי. שימוש בערוץ מדויק הצבעים של העמודות מאפשרים להבחין בקלות בין קבוצות המינים בתוך כל קטגוריה של העדפה מינית.	גרף העוגה מאפשר להמחיש את החלוקה היחסית בין המינים עבור כל ערך של המשתנה העדפה מינית. שימוש בערוצים פחות מדויקים ויותר בזבזניים מבחינת שטח. הצבעים של הפלחים מאפשרים להבחין בקלות בין המינים בתוך כל קטגוריה של העדפה מינית.
אקספרסיביות	השימוש בערוץ הויזואלי מתאים לסוג הנתונים. השימוש בערוצים שנבחרו מתאימים לייצוג של תכונות כמותיות כמו מספר המשתמשים השייכים לכל קומבינציה של מין, נטייה מינית. וכן, המין מיוצג על ידי צבע כיאה לייצוג תכונה קטגוריאלית.	המין מיוצג על ידי צבע כיאה לייצוג תכונה קטגוריאלית. העוגה מציגה את ההבדלים היחסיים בין המינים בתוך כל קבוצה של העדפה מינית על ידי הגדלים של הפלחים.
יתרונות	כל המידע מרוכז בגרף אחד. קל להשוות בין כמות המשתמשים בעזרת גובה העמודות ולהשוות בין המינים עבור כל ערך של העדפה מינית.	הצגה זו מאפשרת השוואה נוחה יותר בהבנה של החלק היחסי של כל מין עבור כל ערך של העדפה מינית.
חסרונות	עשוי שלא להראות ביעילות חלוקה יחסית.	פחות יעיל להשוואת ערכים מוחלטים בין קטגוריות. עבור קטגוריות של מאפיין עם מספר משתמשים דומה, יהיה קשה יותר לראות את ההבדלים. יש צורך בכמה גרפים כדי לענות על השאלה.

בחירה של אחת החלופות למימוש בפרויקט :

בחרנו לממש בחלק האחרון של הפרויקט את החלופה ה-1 (גרף העמודות).

היתרון בחלופה זו הוא שהיא מסכמת את כל המידע בגרף אחד.
בנוסף, היא מאפשרת השוואה ברורה יותר בין הערכים האבסולוטיים (בעזרת המספרים שמופיעים בציר האנכי) וגם ברמת האחוזים (שמופיעים בתוויות על העמודות עצמן)

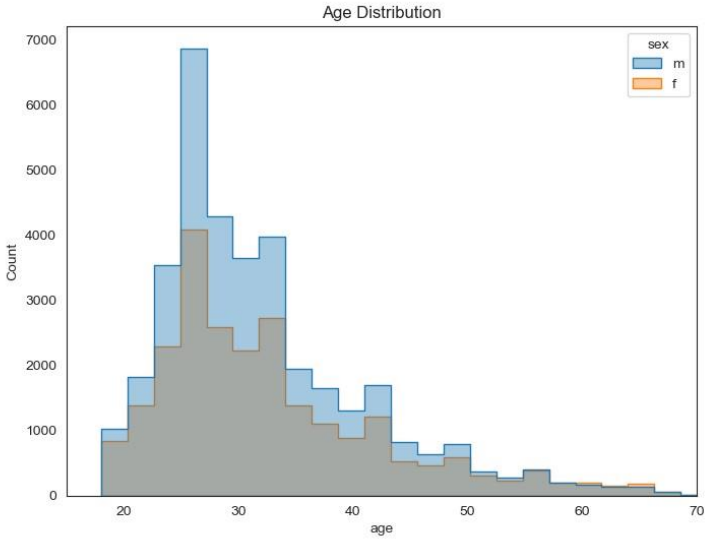
בתרשים הראשון היה עדיף להתחיל מ- Overview: כמה נשים וכמה גברים; כמה בכל אחת משלושת הקטגוריות של הנטייה המינית, כמה מעשנים...

2.כעת, נראה לבדוק את התפלגות הגילאים(משתנה רציף) של משתמשי האפליקציה לפי מין:

מטרת הווזואליזציה: המטרה העיקרית של הסתכלות על התפלגות הגילאים בקרב משתמשי האפליקציה לפי מין. המטרה היא להבין האם יש הבדל בגיל המשתמשים בין המינים, מהו הגיל הנפוץ בקהל היעד והאם יש צרכים לגייס משתמשים מתת קבוצה מסוימת.

חלופה ראשונה: היסטוגרמה

4



משהו לא מסתדר עם הצבעים: מצד אחד נשים מסומנות בכתום, אבל הן בעיקר נראות באפור.

Marks – קווים

כל בין מייצג טווח של גילאים, וגובה הבין מציין את ספירת המשתמשים בטווח הגילאים הזה.

Channels –

מיקום אנכי עם בסיס משותף

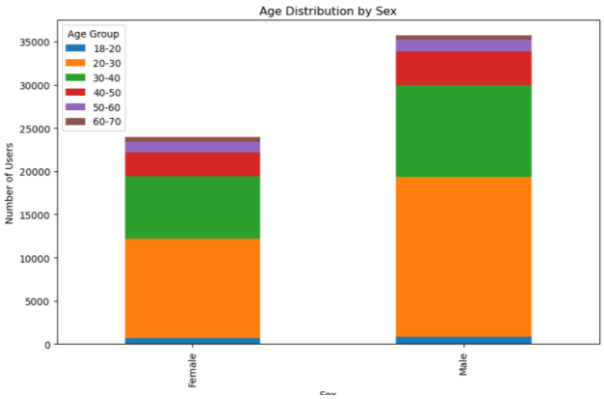
הציר האנכי (גובה העמודה) – מייצג את מספר המשתמשים (לכל טווח ערכים של גילאים).

הפרדה אופקית לפי גיל ממויין

הציר האופקי – בינים של טווחי גילאים.

צבע ההיסטוגרמה – מייצגת את המין .

חלופה 2-



Marks – קווים

כל עמודה מייצגת את אחד המינים – זכר או נקבה.

Channels –

כולל שתי תכונות: ראשית (מין) ומשנית (קבוצת גיל). הראשית והקטגוריה הראשונה מהמשנית ממוקמות על בסיס משותף, שאר הקטגוריות לא על בסיס משותף

הציר האנכי (גובה העמודה) – מייצג את מספר המשתמשים לכל ערך של המאפיין מין .

הציר האופקי - ציר ה-x מייצג את הקטגוריות המושוות (זכר ונקבה).

צבע: צבעים שונים בתוך כל עמודה מייצגים קבוצות גיל שונות, מה שמאפשר בידול ויזואלי של מקטעי גיל בכל קטגוריית מין.

כדאי היה למפות את המיקום הורטיקלי במיקרא למיקום בתרשים

השוואה בין שתי החלופות:

קריטריון	אופציה 1-	אופציה 2 –
אפקטיביות	בהקשר דיוק, הבינים מאפשרים להבחין בכמויות המדויקות של משתמשים בכל קבוצת גיל ובכך לקבל תמונה כמותית מדויקת. ניתן לקבל את התמונה הכמותית המוזכרת לכל מין לפי הצבע של ההיסטוגרמה ובכך להבחין בין הקבוצות השונות- באמצעות יצירת 2 היסטוגרמות	הגרף מאפשר לראות את הערך הכולל של המשתמשים בכל מין בצורה מדויקת, כמו גם את החלוקה הפנימית של תתי-קטגוריות של גילאים בתוך כל עמודה לפי הצבעים. זה

	אחת על גבי השנייה בצבעים שונים, ניתן להשוות בניהן בצורה נוחה.	מאפשר השוואה קלה בין חלקים שונים בתוך אותו מין.
אקספרסיביות	השימוש בערוץ הויזואלי מתאים לסוג הנתונים. השימוש בערוצים שנבחרו מתאימים לייצוג של תכונות כמותיות כמו מספר המשתמשים השייכים לכל טווח של גילאים . המין מיוצג על ידי צבע כיאה לייצוג תכונה קטגוריאלית - הצבעים של ההיסטוגרמות מאפשרים להבחין בין ההתפלגויות של שני המינים.	תתי החלקים שיצרנו לכל קבוצת גיל מיוצגים על ידי הצבעים השונים ובכך באים לייצג תכונה עם מספר ערכים. כמו כן, השימוש בעמודות מאפשר לראות את הערך הכמותי הכולל של המשתמשים בכל מין בצורה מדויקת.
יתרונות	התרשים מאפשר להשוות בין התפלגות הגילאים של שני המינים. ניתן לראות בקלות את הכמויות המדויקות של משתמשים בכל קבוצת גיל ובכך לקבל תמונה כמותית מדויקת. ההיסטוגרמה מאפשרת זיהוי דפוסים וחריגים בהתפלגות הגילאים של כל מין, כמו גילאים שבהם יש פערים גדולים בין המינים או גילאים שבהם יש מספר דומה של נציגים. ניתן להשוות בין מרכז ההתפלגות של כל גיל ולמצוא דמיון או שוני.	ניתן לראות גם את הסכום הכולל של כל קטגוריה וגם את תרומתם היחסית של כל תת-חלק לסכום הכולל. זה מקל על הבנת הדינמיקה הפנימית של כל קטגוריה - ניתן לראות גם את החלוקה הפנימית של הגילאים בתוך כל עמודה לפי מין. זה מאפשר השוואה קלה בין הגילאים השונים עבור כל מין.
חסרונות	היסטוגרמה מציגה את הנתונים לפי בינים, ולכן המידע הפרטני אודות כל נקודת נתונים אובד. לדוגמה, לא ניתן לדעת את הגיל המדויק של המשתמשים, אלא רק את הקבוצה אליה הם משתייכים.	כאשר יש הרבה תתי-חלקים בתוך כל עמודה, יכול להיות קשה להשוות את הגודל של כל תת-חלק בין עמודות שונות - במיוחד כאשר הם אינם מתחילים מאותה נקודת ייחוס. גרף כזה עם הרבה תתי-חלקים יכול להיות מסובך לקריאה ולהבנה. הרבה צבעים ושכבות עלולים לגרום לבלבול ולירידה ביכולת לבצע השוואות מהירות.

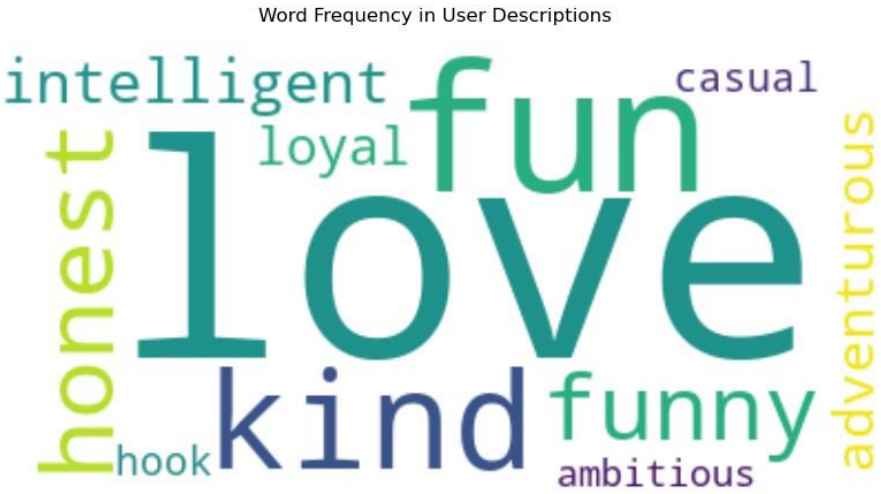
בחירה של אחת החלופות למימוש בפרויקט :

אופציה 1 (היסטוגרמות) נבחרה מכיוון שהיא מאפשרת לנו להציג את ההתפלגות הכמותית המדויקת של המשתמשים לפי קבוצות גיל ומין בצורה ברורה ונוחה. היכולת לראות את הכמויות המדויקות בכל קבוצת גיל וההשוואה הקלה בין ההתפלגויות של שני המינים מציעה תמונה כמותית מדויקת ומאפשרת השוואה וזיהוי דפוסים בנתונים בצורה אפקטיבית.

3.כעת נרצה לענות על השאלה **מהן התכונות הערכים וסוגי הקשרים שמשתמשים מחפשים בקשרים דרך אפליקציית היכרויות?**

מטרת הוויזואליזציה : החשיבות של ניתוח המילים בפרופילים של משתמשים באפליקציית היכרויות מאפשרת הבנה עמוקה יותר של רצונותיהם ותחומי העניין שלהם. באמצעות זיהוי מילות מפתח כמו "כנות", "מצחיק", "אינטליגנט", וכדומה, אנו יכולים להבין את המאפיינים שהמשתמשים מעריכים באנשים או בקשרים שהם מחפשים. זה מאפשר למפתחי האפליקציה לייעוץ ולשפר את החוויה הכללית, להתאים את ההצעות ולהגביר את הסיכויים למצוא התאמה מתאימה בהתאם לציפיות ולדרישות המשתמשים.

חלופה 1:



בקורס לא כללנו מילים כסימנים. אני חושב שהסוג הזה של ויזואליזציה קצת חורג מהתחום שלנו.

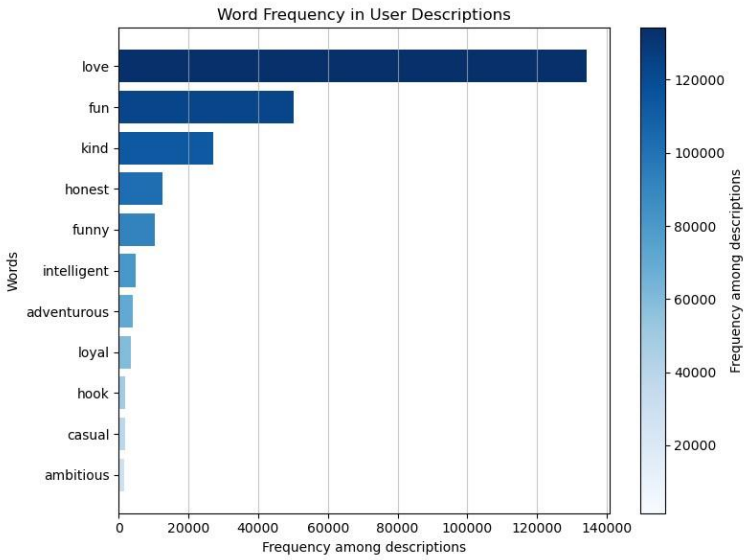
Marks –

מילים: המילים לקוחות מתוך תיאור הפרופיל של המשתמש ומענה שלו על שאלות באפליקציה.

Channels –

גודל: הגודל של כל מילה הוא פרופורציונלי לתדירות ההופעה שלה בפרופילים של המשתמשים. כלומר, ככל שהמילה גדולה יותר כך היא הופיעה יותר פעמים בתיאור הפרופיל של המשתמשים באפליקציה.

חלופה 2:



Marks- קווים

ברים - כל בר מייצג מילה ספציפית מתוך רשימה, ואורך הבר מתאים לתדירות של מילה זו בתיאורי המשתמש באפליקציה.

Channels –

גוון: הגוון של כל פס מייצג את התדירות של כל מילה בפרופילים של המשתמשים בנתונים.

גוונים כהים יותר מעידים על מילים שכיחות יותר, שהופיעו בתדירות גבוהה יותר בנתונים ולהפך.

האם יש סיבה לקידוד כפול של מספר המילים בכל קטגוריה (מיקום וצבע)?

הציר האנכי: בתרשים מייצג את המילים מתוך הרשימה המסוימת שנבחרו.

הציר האפקי: המיקום האופקי והאורך של כל פס מייצגים את התדירות של המילה המתאימה בתיאורי הפרופילים של המשתמשים.

השוואה בין שתי החלופות:

קריטריון	אופציה 1-	אופציה 2 –
אפקטיביות	פשוט להבין תדירויות של מילים ספציפיות באמצעות גודלן במרחב אך ללא יכולת מדויקת לדעת את התדירות של כל מילה.	מאפשר מדידה והערכה מדויקת של כמו הפעמים שהמילה הופיעה בפרופילים של משתמשים. כאן באה לידי ביטוי יכולת גבוהה לדייק בתדירויות וכן להבחין בשוני בין התדירויות של מילים שונות על ידי אורך הבר וכן הגוון של הצבע.
אקספרסיביות	מילים גדולות הן בולטות לעין של המשתמש, וקל להבחין בהן במהירות.	השימוש בערוץ הוויזואלי מתאים לסוג הנתונים. השימוש בערוצים שנבחרו מתאימים לייצוג של תכונות כמותיות כמו מספר המופעים שהופיעה כל מילה. בנוסף, הגוון של הצבע מתאים לייצוג של תכונה כמותית.
יתרונות	פשוט להבין תדירויות של מילים ספציפיות באמצעות גודלן במרחב. מאפשר הצגה ויזואלית יפה וקליטה של מילים בולטות. זו אולי דוגמה טובה לאסתטיקה שקודמת לפונקציונליות - למדנו שזה לא מומלץ בתחום	מאפשר מדידה והערכה מדויקת של כמו הפעמים שהמילה הופיעה בפרופילים של משתמשים. קל לבצע השוואות מדויקות בין מילים שונות לפי גובה העמודות וגוון הצבע.
חסרונות	לא ניתן להבדיל בין החשיבות של מילים שהופיעו מספר פעמים דומה בנתונים, גודל המילים יהיה מאוד קרוב ולא בולט לעין.	כאשר נרצה לבחון את החשיבות של מספר רב של מילים הגרף עשוי להראות עמוס ויזואלית.

בחירה של אחת החלופות למימוש בפרויקט :

לכל אחד מהגרפים יש יתרונות מהותיים שונים ולכן במקרה זה הבחירה תלויה במטרה שהגדרנו לעצמנו בשאלת המחקר. המענה העיקרי שרצינו שהוויזואליזציה תשקף הוא מהם מאפייני הקשר שהמשתמשים באפליקציה מחפשים. רצינו לתת חשיבות לכאלו שהם מבוקשים יותר מאחרים ולהבליט אותם מבחינת **סדרי הגודל**, אך לא היה ברצוננו לתת מענה שמדרג אותם בצורה מסודרת אחד אחריי השני.

לכן, החלטנו לבחור בוויזואליזציה הראשונה מפני שאנחנו חושבות שהיא מיוחדת ומעניינת יותר את הצופים בה וכן מעבירה את הרעיון שרצינו בצורה מספקת.

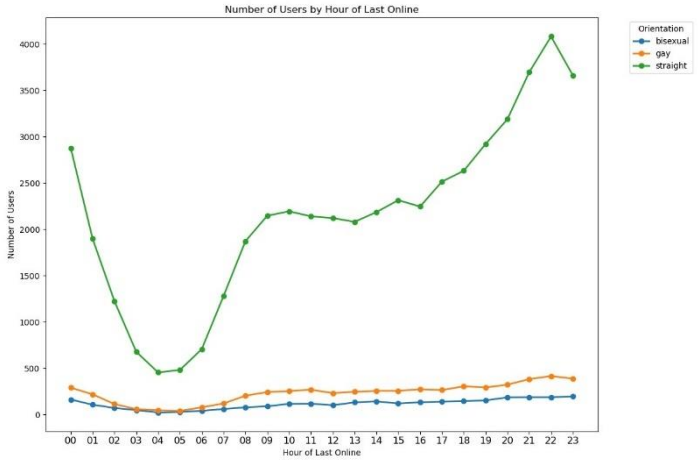
4. זיהוי מגמות התחברות לאפליקציה לאורך היום – עד איזה שעות ביום משתמשים נוטים להשתמש באפליקציה לפי נטייתם המינית?

מטרת הוויזואליזציה : באמצעות התחשבות בשעת ההתחברות האחרונה, ניתן לבנות פרופילים של משתמשים שונים.

בהבנת השעות המובילות לשימוש, ניתן לפתח אפשרויות לשדרוג ולשיפור חווית המשתמש באפליקציה. לדוגמה, ניתן להגדיר אפשרויות שימוש שונות, פרסומות, או פעולות שיווק בהתאם לזמני השימוש הפופולריים.

בהתבסס על המגמות בשעות ההתחברות האחרונה, ניתן לייעוץ על אופטימיזציה של משאבים, כמו שרתים או מערכות אחסון, כדי להתאים אותם לזמני השימוש המרכזיים.

חלופה 1- line plot



–Marks

נקודות וקווים
קווים - מציגים מגמות ושונות בספירת המשתמשים על פני שעות.

סמן נקודות ('ס'): הסמן ('ס') מאפשר לראות ספירות שעות מדויקות.
להפך - קודם יש את הנקודה, ואח"כ מחברים נקודות סמוכות באמצעות קו.

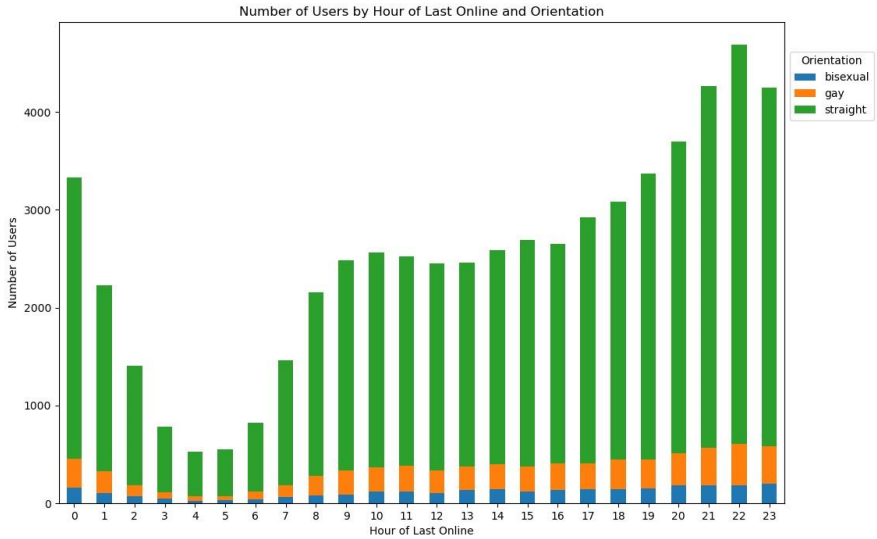
– Channels

ציר אופקי (שעה אחרונה באינטרנט): ציר זה מראה את ערכי השעות שהמשתמשים התחברו בהן לאחרונה.

ציר אנכי (מספר משתמשים): ערוץ זה מקודד את ספירת המשתמשים, הממוקמת אנכית על סמך מספר המשתמשים בכל שעה.

צבע (כיוון): כל קו מובחן לפי צבע. הצבע עוזר להבחין בין קבוצות שונות של משתמשים לפי נטייתם המינית.

חלופה 2- stacked plot



– Marks

ברים - כל בר מייצג את ספירת המשתמשים עבור שעה ספציפית של התחברות אחרונה לאפליקציה במהלך היום, מפולח לפי נטייה מינית.

– Channels

הציר האופקי (ציר x): מתאר את השעה האחרונה שבה המשתמש התחבר לאפליקציה. מיקומו של כל בר לאורך ציר ה-x מתאים לשעה ספציפית ביום שבה המשתמשים היו מקוונים לאחרונה.

הציר האנכי (ציר y): מתאר את מספר המשתמשים שהתחברו לאחרונה באותה שעה.

צֶבַע: נטיות מיניות שונות נבדלות בצבעים שונים. כל קטגוריה של נטייה מינית מיוצגת על ידי צבע שונה.

השוואה בין שתי החלופות -

קריטריון	אופציה 1-	אופציה 2 –
אפקטיביות	גרף קו מותאם להצגה של נתונים לאורך זמן, מאפשר לראות מגמות וטרנדים בבירור. קו מגמה לאורך זמן מאפשר לראות בצורה בולטות דפוסים חריגים.	הגרף מאפשר לראות את הערך הכולל של המשתמשים המחוברים בשעה מסוימת, כמו גם את החלוקה הפנימית של תתי-קטגוריות של נטייה מינית בתוך כל עמודה לפי הצבעים. זה מאפשר השוואה קלה בין חלקים שונים בתוך אותה נקודת זמן (שעה).
אקספרסיביות	יצירת קו לכל קבוצה של נטייה מינית (לפי צבע) מאפשרת להבין את הטרנד עבורה בצורה ברורה ופשוטה. הצבע תואם את ההבחנה בין משתנים קטגוריאליים. הבחירה בגרף מגמה מתאים להראות את השינוי במספר המשתמשים המחוברים בשעות היממה (משתנה כמותי) השימוש בערוץ הוויזואלי הנ"ל מתאים להצגה של נתונים לאורך זמן.	תתי החלקים שיצרנו לכל קבוצה של נטייה מינית מיוצגים על ידי הצבעים השונים ובכך באים לייצג תכונה קטגוריאלית עם מספר ערכים. כמו כן, השימוש בעמודות מאפשר לראות את הערך הכמותי הכולל של המשתמשים שהתחברו לאחרונה בכל שעה ביממה בצורה מדויקת.
יתרונות	גרף קו מותאם יותר להצגה של נתונים לאורך זמן, מציג את השינויים בנתונים, מאפשר לראות מגמות וטרנדים בבירור.	גרף מאפשר לראות בצורה מדויקת אילו שעות ביום נפוצות

	יצירת קו לכל קבוצה של נטייה מינית מאפשר להבין את הטרנד עבורה בצורה ברורה ופשוטה. מאפשר להשוות בין המגמות בקרב כל קבוצה בצורה נוחה.	ביותר להתחברות לאפליקציה בקרב כל הקבוצות יחד.
חסרונות	מדגיש את המגמה של כל קבוצה בפני עצמה ופחות את המגמה הכללית בקרב הנתונים.	הבנה מורכבת ואיטית יותר של הטרנד של כל קבוצה בנפרד.

בחירה של אחת החלופות למימוש בפרויקט :

בחרנו בחלופה 1 כי היא מאפשרת לראות מגמות וטרנדים לאורך זמן בצורה ברורה. גרף הקו מתאים במיוחד להצגה של שינויים בנתונים ומאפשר הבנה ברורה של הטרנדים עבור כל קבוצה של נטייה מינית, תוך השוואה נוחה בין המגמות של הקבוצות השונות.

הסבר על העיצוב שנבחר ועל יישומו:

<u>עיבוד מוקדם של הנתונים:</u>

לטובת מימוש של הוויזואליזציה השלישית, בה רצינו לענות על השאלה:

מהן התכונות הערכים וסוגי הקשרים שמשתמשים מחפשים בקשרים דרך אפליקציית היכרויות?

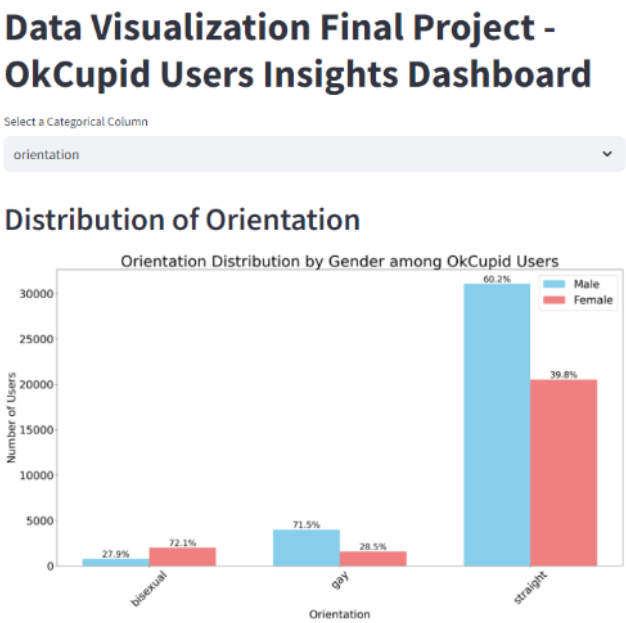
יצרנו עמודה חדשה בנתונים אשר מאחדת עבור כל משתמש את כל המלל החופשי והמענה שלו לשאלות (עמודות Essay0-Essay9) , מה שאפשר לנו לבנות את הוויזואליזציה בצורה נוחה יותר שמסתמכת על עמודה זו.

לא שינינו את מימוש החלופות שבחרנו, אך במהלך בניית הדשבורד חשבנו איך נוכל להוסיף דינאמיות לכל וויזואליזציה על מנת להעשיר את המידע שהמשתמש יוכל לקבל ממנה.

חבל שאין מבוא/הסברים בדשבורד באופן כללי ולפני כל תרשים באופן ספציפי

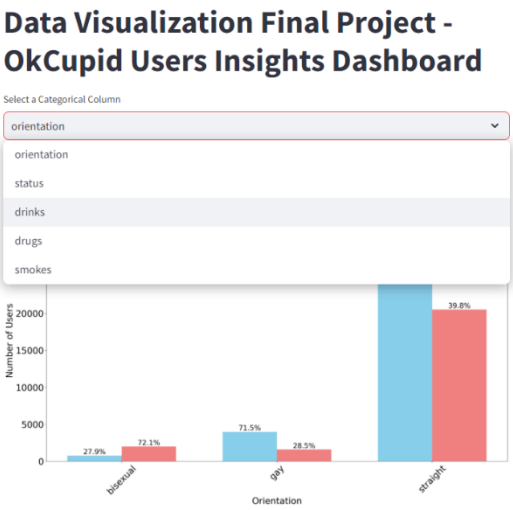
<u>צילומי מסך מתוך הוויזואליזציה:</u>

<u>וויזואליזציה 1:</u> הצגת ההתפלגות של פיצ'רים קטגוריאליים שנבחרים שונים לפי מין

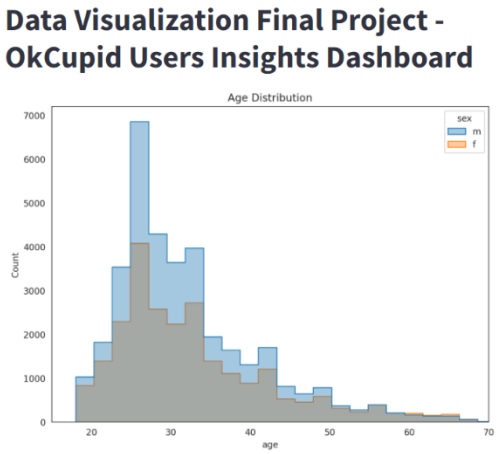


כאן במיוחד יש מקום להערה שהאחוזים מתייחסים לחלק מתוך הקטגוריה המשנית ולא מתוך הקטגוריה הראשית (גברים/נשים)

ניתן לבחור באופן דינאמי את **הפיצ'ר המבוקש לפיו נציג את ההתפלגות:**

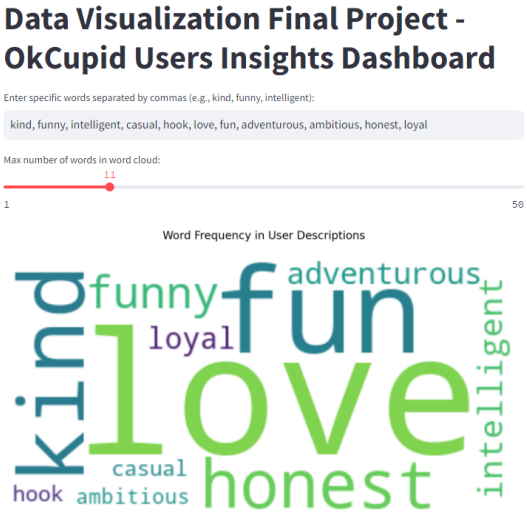


<u>ויזואליזציה 2:</u> הצגת התפלגות הגילאים של המשתמשים לפי מין



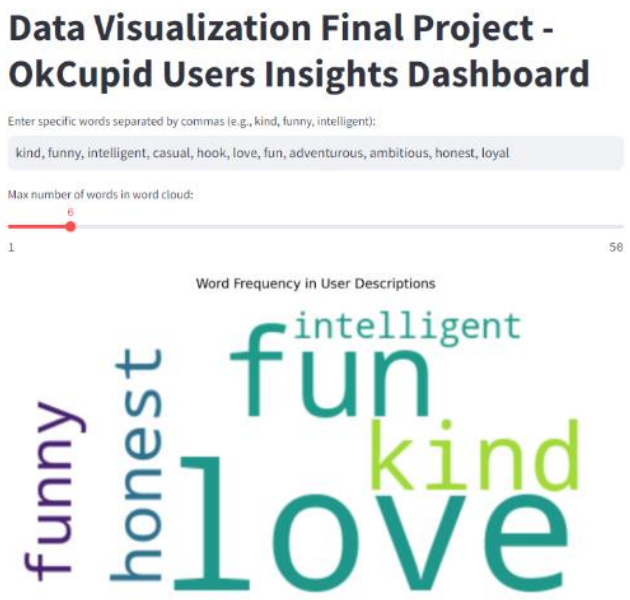
הסבר?

<u>ויזואליזציה 3:</u> הצגת התכונות וסוגי הקשרים שמשתמשים מחפשים באפליקציה

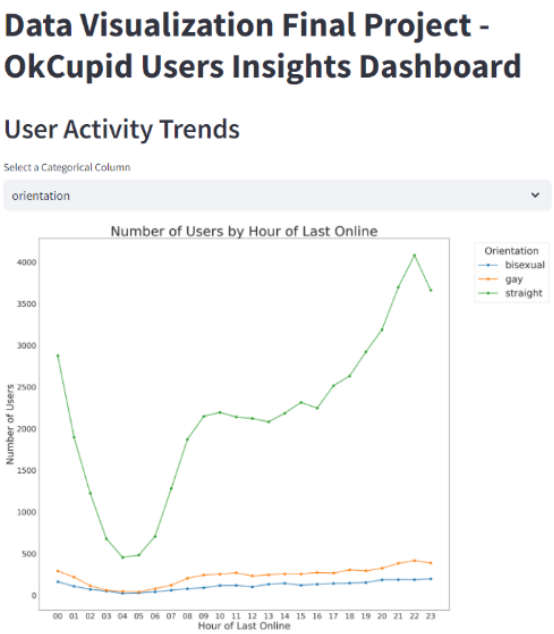


ניתן להגדיר את המילים שהמשתמש רוצה לבדוק ולחפש (להאריך או לקצר את הרשימה) , וכן לבחור את כמות המילים הכי נפוצות שנרצה להציג בסופו של דבר בתצוגה.

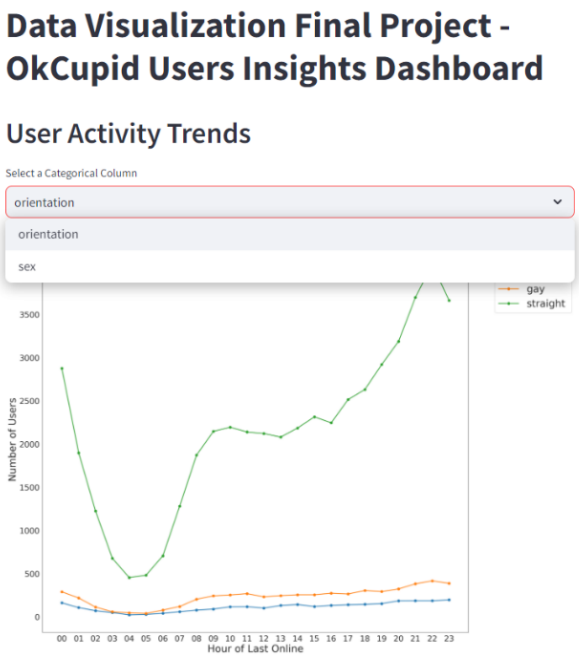
לדוגמה: בחרנו את אותן מילים כמו בדוגמה הקודמת, אך רצינו לראות רק את ששת המילים הכי נפוצות ואת יחסי הסדר בניהן:



<u>ויזואליזציה 4:</u> זיהוי מגמות התחברות לאפליקציה לאורך היום לפי הפיצ'ר שנבחר



ניתן לבחור באופן דינאמי את **הפיצ'ר המבוקש לפיו נציג את ההתפלגות:**



כפי שאפשר לראות, כאשר יש הבדל גדול בין הקטגוריות השונות, אז השינויים בקטגוריות הפחות שכיחות הולכים לאיבוד. אפשר לנסות לטפל בזה באופן אינטראקטיבי, למשל ע"י אופציה להוריד את הקטגוריה הדומיננטית מהתרשים.

<u>תהליך כתיבת הקוד:</u>

במהלך תהליך כתיבת הקוד, לא השתמשנו בקוד קיים אלא בנינו את כל הקוד במהלך העבודה על הפרויקט.

כדי לממש את החלק של העיצובים החלופיים יצרנו מחברת פייתון שבונה את הויזואליזציות בעזרת שימוש בפונקציונליות המוצעת על ידי שתי ספריות עיקריות: matplotlib,seaborn.

בנוסף השתמשנו בספריית wordcloud שמאפשרת להציג תדירות של מילים בטקסט בצורה גרפית. גרפים אלו הם דרך ויזואלית להדגיש מילים חשובות או שכיחות בטקסט נתון, כאשר גודל המילה מציין את תדירותה.

לאחר שבנינו את המחברת המתוארת ובחרנו את החלופות המועדפות עלינו למימוש בפועל, בנינו קוד נוסף שבונה את הדשבורד עם הויזואליזציות שנבחרו בעזרת streamlit.