פרוייקט גמר – ויזואליזציה

World Happiness data set

מגיש:

דניאל קוזנצוב 318856648

ויזואליזציה: https://dandan.streamlit.app/

קריאה מהנה 😊😊😊

מבוא:

ברוכים הבאים לוויזואליזציה עבור דאטהסט בין הפופולריים היותר בעולם: World Happiness dataset . דאטהסט זה נחשב מאוד מעניין כיוון שהוא עוסק במדד השמחה בחלקים שונים בעולם לפי מדד הLadder ובנוסף לכך מכיל עוד כמה פרמטרים מאוד מעניים אשר עלולים להשפיע על השמחה עבור מדינה.

בוויזואליזציה זו ננסה לענות על כמה שאלות מאוד מעניינות שנראה כי עדיין לא נענו בעזרת ויזולאיזציות. השאלות העיקריות שלי נובעות מהניסיון להבין לעומק את המידע אשר קיים שם ביחס לשנים ולאזורים.

השאלות הינם: 1) מה המדינה הכי שמחה לאורך השנים? מה הממוצע של מדד השמחה בעולם?

אך ישנם שאלות שידרשו צלילה לעומק בעזרת הויזואליציה:

2) מה בעצם משפיע על השמחה? על מנת שנוכל לענות על השאלה הזו נרצה לצלול אף יותר לעומק

- איך המדדים השתנו עם השנים?
- איך המדדים משפיעים באזורים שונים בעולם?
- האם קיימים קשרים בין המדדים השונים?

3) עבור המדדים השונים מי מהמדינות המדורגות בשיא?

4) האם יש הבדלים מהותיים בשמחה בין חלקים שונים בעולם?

על כל השאלות האלה נענה בעזרת ה World Happiness dataset מKaggle

הדוח נחשב לאמין כיוון שהינו מגיע מארגון הSustainable Development Solutions Network בארצות הברית. והאיחוד מידע יבוצע על ידנו.

היקף הנותנים הינו עצום בעל 936 רשומות עבור מדינות שונות בין השנים 2015 עד 2020. והאיחוד של הדאטה יבוצע על ידי סקריפט בפייתון שיצורף בסוף.

נתונים :

הנתונים אשר קיים אצלנו בדאטהסט הינם:

מדד	מספר רשומות	הסבר	משמעות למטלה
Country name	934	השם של המדינה	בעזרת הנתון הבא יבוצעו האגרגציות והחלוקות, מדד חשוב ביותר
Regional indicator	936	מחלק את המדינות לאזורים	כיוון שמעניין אותנו מאוד ההבדלים באזורים שונים נוסיף מידע זה לדאטה ונעשה בו שימוש.
Year	936	השנה שבה נאסף המידע	כיוון שמעניין אותנו מאוד ההבדלים בשנים שונות מידע זה מאוד חשוב לנו ועל כן ביצענו את האיחוד בין הדאטהסטים השונים.
Ladder score	936	תוצר מקומי גולמי, התפוקה הכלכלית של מדינה מחולקת ליחס האנשים בה	מדד זה הינו מאוד חשוב לנו כיוון שנרצה לבדוק כיצד הוא משפיע על השמחה והאם הוא משפיע על מדדים אחרים
Social support	936	משקף את הרשתות תמיכה שיש לאדם במדינה. כגון משפחה, חברים ועוד אשר עלולים לתמוך באדם נפשית	מדד זה גם מעניין אותנו כיוון שאנו אכן חוששים שיש לו השפעה על שמחה ונרצה להציג אותו בגרפים שלנו
Healthy life expectancy	936	משקף את מספרים השנים הממוצע שאדם יחיה במדינה. לפי פקטורים של סגנון חיה ורפואה במדינה	מדד זה בנוסף הוא אחד מהמדדים שאנו מצפים שמאוד ישפיע על שמחה ועל כן נרצה לתעל אותו לטובת הויזואליזציה
Freedom to make life choices	936	מעריך את הרמה של חופש ואטונומיה עבור אדם במדינה. מציג עד כמה אדם חופשי, ובעל שליטה על הבחירות בחייו	כמובן שמדד זה הוא מאוד משמעותי עבור איכות חיים לכן נרצה גם לבדוק את ההשפעה שלו ולהציג מדינות מובילות בתחום זה
Generosity	936	המדד מעריך את הנכונות לעזור לאדם אחר עבור מדינה, מבוסס על סקר תרומה לקהילה	מדד זה מאוד מעניין אותנו גם כן כי ידוע לנו שתרומה לזולת עוזרת לתחושת הסיפוק העצמי
Perceptions of corruption	936	המדד מעריך את רמת השחיטות במדינה	ידוע כי שחיטות במדינה עלול להוביל לאסונות לכן ניקח מדד זה בחשבון
Latitude (lat) and Longitude (long):	936	מידע גאוגרפי למדינות לשימוד בויזואליזציות	-

הסבר על העיצוב שנבחר:

תחילה נראה כי ישנה כמו סטטיסטיקה מהירה עבור השמחה הממוצעת בעולם, תוחלת חיים, המדינה השמחה ביותר, ציון השמחה בה והתוחלת חיים בה. אמנם זה אינו עונה על הקריטריונים של מזנמר במונחי marks and channels אך זה מידע מאוד חשוב לעניות דעתי שיש לשים בראשית הדאשבורד על מנת שהמשתמש יוכל לקבל סטטיסטיקה מהירה עבור הדברים שהוא בחר.

אך למרות זאת מבחינת האפקטיביות והאקספרסיביות הצגת הנתונים בצורה זו יכול להיות טובה כיוון שמבחינת האקספרסיביות, אנו מקבלים מידע מהיר עבור השמחה הכללית בעולם, כך שאני מקבלים הבנה טובה של השמחה העולמית. בנוסף לכך אנו מקבלים הבנה טובה עבור הכמות חיים ממוצעת עם כל ההשלכות של זה. ויותר מכך ההבנה שאנו מקבלים על השמחה עבור המדינה והפרטים עליה כבר מוסיפים הבנה עם ניואנסים על הדאטה. לא מבחינת אקספרסיביות יכול להיות שזה רעיון טוב. מבחינת אפקטיביות, המידע שהוצג נותן הבנה מהירה על נקודות מפתח בדאטה והשימוש בערכים נומרים מחזק את האפקטיביות במקרה זה.

לכן במונחים של אקספרסיביות, המידע המסופק מכסה היבטים מרובים של אושר, כולל ממוצעים גלובליים ודוגמאות מדינות ספציפיות, מה שמגביר את כושר הביטוי שלו. מבחינת אפקטיביות, המידע מוצג בצורה ברורה ותמציתית, מה שמקל על ההבנה והעיסוק בנתונים.

אך כמובן שיש לציין שזה לא המדדים המקובלים של מנזנר ולכן תיארתי אותם לפי דעתי.

לאחר מכן השימוש בHeatmap אזורי, מבחינת marks מפת החום משתמשת בצורות של המדינה (מרובעים או מלבנים) על מנת לייצג כל מדינה. כלומר מבחינת marks ישנו שימוש בareas . במונחים של channels , אנחנו משתמשים בcolor על מנת לייצג את המדינות השמחות יותר או השמחות פחות.
במונחים של אפקטיביות, המפת חום נותנת לנו הצגה ברור ומתומצתת של הדירוג אושר עבור אזורים שונים. ובכך היא נותנת למשתמש בקלות להבין הבדלים ברמת האזורים עבור הרמת אושר.
בסקייל אשר הגדירה מנזנר ניתן לראות כי המפה עומד תחת area, color saturation ועל כן אינה מאוד אפקטיבית אך יש פה גם עניין של הכרות עם מפת העולם שקיימת עבור כל אחד ועל כן ניתן בקלות לדעת על איזה אזור מדובר ומדינה, כלומר גם עבור מדינה קיים identity channel של spatial region . ועל כן לדעתי הגרף הינו מאוד אפקטיבי על מנת להסביר את מה שנדרש עבור חלק זה
בנוסף לכך גם הכניסה לשנים ואזורים גורמת לכך להיות מאוד אפקטיבי בהבנה לעומק של הדאטה בקלות.

במונחים של אקספרסיביות, מפת החום מתקשרת בצורה טובה את הרמת אשור עבור אזורים שונים בזמנים שונים, השימוש בColor saturation אף מדגישה כך. והחלוקה ותצוגה לאזור אף נותנת הבנה של ההתפלגות המרחבית.
לפי ההגדרות של מנזנר:

1) דיוק: המפת חום יכולה לייצג את הגירוי (Stimulus) על ידי הsaturation כלומר הוא יוצר יותר גירוי ממה שאמור באמת להיות. אך עבור השאלה שנשאלה (4) אין משהו יותר טוב לדעתי אשר יכול להציג את השאלה הזו
2) יכולת הפליה: המפת חום יעילה בהבחנה בין מדינות שונות ורמת השמחה בהן. כל מדינה משורטטת לפי מה שמוכר לנו (ואף משתנה לקטגוריה) , מה שמקל על הבחנה בין מדינות שונות. אך ההפרדה מבחינת הרמת שמחה נהיית לא פשוט כיוון שהצבע לא שונה בצורה מהותית ועל כן נוסף הפיצ'ר שבעזרת ריחוף העכבר מעל המדינה נקבל את שמה ורמת השמחה שבה
3) יכולת הפרדה: מהיות המפת חום מציגה את כל הנתונים עבור שנה מסוימת ואזור מסויים עם הפרדות בין המדינות, ניתן להגיד שמבחינת מדד זה הבחירת במפת חום היא מעולה. אך מפת החום סובלת מכך שהמדינות חופפות אחת לשניה ואי אפשר להגיד שאכן קיימת פה הפרדה ועל כן גם במדד זה לוקה בחסר. למרות הצגת קווים ברורים בין המדינות וצבעים שונים.

4) בצבוץ: מהיות המפת חום מציגה את המדינות כמו שהם עם הצבע עבור השמחה, אין פה יכולת לגרום למדד זה להיות טוב ועל כן גם בתחום זה לוקה בחסר.
5) הקבצה: השימוש במדינות כמו שאנו מכירים אותם על המפה עובד במפת חום שלנו בצורה נפלאה ואכן ניתן לראות את ההבדלים בין המדינות השונות. קל להבחין במדינה ולקבל מידע עבור כולה.

לפי המדדים אומנם המפת חום אינה הצורה האידאלית על מנת להציג את השאלה הזאת, אך לא מצאתי אפשרות אחרת שתסביר את השאלה (4) כמו שאני רוצה ועל כן אעדיף להישאר איתה.

לאחר מכן מתבצע פילטר לפי הפיצ'רים שמעניינים אותנו ונקבל שתיי גרפים שיענו על השאלות שלנו.

תחילה נדון בscatterplot :

מבחינת marks יש שימוש בmarks מסוג points . כאשר אנו צוללים יותר עמוק נראה כי ישנם הרבה visual channels שבהם נעשה שימוש: 1) position – נעשה שימוש בboth X-axis and Y-axis כך שנוכל לקבל את המימד בו נמצאים הנקודות ונבין את הקשר בין הפיצ'רים 2) ישנו שימוש Size-area שהוא תלוי ברמת השמחה עבור מדינה כך שעדיין נוכל לקשר את הפיצ'רים לרמת שמחה כאשר אנו בודקים פיצ'ר כנגד פיצ'ר
3) ישנו שימוש בcolor עבור האזורים השונים שניתן לבחור בצד שמאל של הדאשבורד.
כלומר הscatterplot שיצרנו עונה בצורה מעולה על השאלות שאנו שואלים ועומד בצורה מושלמת בתנאים של מנזנר עבור ויזואליזציה.

במונחים של אפקטיביות, הscatter plot מציג באפקטיביות את היחס בין התכונות שנבחרו לבין רמת השמחה. כלומר מאפשר למשתמשים לנתח חזותית את נקודות הנתונים ולזהות דפוסים, אשכולות ומגמות. השימוש בצבעים שונים לייצוג אינדיקטורים אזוריים שונים משפר את האפקטיביות על ידי מתן מידע קטגורי נוסף. וכיוון שיש פה שימוש בsize area וposition גם ניתן להעריך את זה גבוה בסולם האפקטיביות.

במונחים של אקספרסיביות, הscatter plot מעבירה ביעילות את התפלגות נקודות הנתונים על פני צירי x ו-y, כמו גם את הגדלים היחסיים של הסמנים המייצגים את ציוני השמחה. פונקציונליות הרחף, הצגת שם המדינה בעת ריחוף מעל נקודת נתונים, מוסיפה יכולת ביטוי על ידי מתן מידע ספציפי על נקודות נתונים בודדות.

בהיכנס למונחים שבהם מנזנר הגדירה נגיד כי:
1) דיוק: הscatter plot יכולה לייצג את הגירוי (Stimulus) על ידי תיאור מדויק של הקשר בין שני משתנים רציפים על הגרף. יותר מכך השימוש בצבעים ובשטח לפי הרמת שמחה אף מוסיף למדד זה על פי מנזנר
2) יכולת הפליה: הscatter plot יעילה בהבחנה בין רמות או קטגוריות שונות של תכונת נתונים. כל נקודת נתונים משורטטת כסמל נבדל (ואף סמל משתנה לקטגוריה), מה שמקל על הבחנה בין נקודות בודדות זו מזו. מהיות כך שהscatter plot מציע visual channel דיי מצומצם לנתונים שאני רוצה להציג הוא לא משולם בקטגוריה הזו, אך הגרפים האחרים שיכלו להציג זאת מאוד מורכבים עבור המשתמש
3) יכולת הפרדה: מהיות הscatter plot מציג את כל הנתונים עבור שנה מסוימת הוא סובל גם בחלק זה ולוקה בחסר אך על מנת לטפל בכך אני משתמש בסינון של שנה ואזור כך שהוא יותר מצומצם ומופרד לדרישות שלנו.
4) בצבוץ: מהיות כך שאנו משתמשים בצורות שונות עבור כל אזור ואף צבע שונה עבור כל אזור, המדד הזה פועל מצוין אצלנו בוויזואליזציה
5) הקבצה: השימוש בצורות שונות וצבעים שונים עובד בגרף שלנו בצורה נפלאה ואכן ניתן לראות את ההבדלים בין הקבוצות השונות

כלומר על פי כל המדדים וההסבר הוויזואליזציה לטעמי מעולה לשאלה שאני שואל (2) של קשר בין הפיצ'רים ובכללי הקשר למדד השמחה.

ניתוח Bar plot :
מבחינת marks יש שימוש בmarks מסוג lines . כאשר אנו צוללים יותר עמוק נראה כי ישנם הרבה visual channels שבהם נעשה שימוש: 1) position – נעשה שימוש בY-axis כך שנוכל לקבל את המספר של המדד בצורה יותר איכותית 2) ישנו שימוש Size-Length שהוא תלוי בפיצ'ר אשר מבחר, בעצם נקבל מעין צורה של barplot עולה וכך נקבל דירוג עבור המדינות (ונוכל לקבל דירוג מדויק עבור כל בר בעזרת ריחוף מעל הבר עם העכבר)
3)ישנו שימוש בcolor עבור האזור שבו המדינה נמצאת. נותן מעין הרגשה של קלסטרינג ומראה הבדלים משמעותיים בין אזורים. (ואף ניתן לסנן בעזרת הslider a ladder score כדי לקבל הדגמה יותר מדויקת)

כאשר נבחר לנתח בעזרת מונחי אפקטיביות נראה כי אנחנו משתמשים באורך על מימד אחד וגם צבעים ועל כן נוכל להגיד שchannel שלנו הוא אכן מתאים ומאד אפקטיבי לתיאור המידע, בנוסף לכך עבור כל אזור הבר נצבע בצבע ייחודי ובכך נוכל להגיד שהchannel הוא מאוד אפקטיבי בהעברת המידע שלו. ואכן כך נראה כי ניתן להבין המון לגביי הפיצ'ר בין המדינות לפי האזורים. ואף אפשר לעשות זום אין לתוך הגרף על מנת לקבל המחשה יותר טובה.

כאשר נבחר לנתח בעזרת מונחי אקספרסיביות נראה כי:

1) דיוק: מבחינת דיוק כמובן שבר פלוט הוא בין הגרפים האידיאלים למטלות אלה, גם על פי מנזנר הוא מאוד חזק עבור ההבנה שלנו ונחשב לערוץ ויזואלי מעולה.
2) יכולת הפליה: מבחינת יכולת הפליה גם כן גרף מסוג בר פלוט הוא מעולה כיוון שניתן בקלות להבין את ההבדלים בbins בין המדינות השונות (וניתן אף לעשות זום אין), עקב כך שמדד זה מושג על ידי כך שניתן להבחין בהבדלים בעזרת האורך שזו גם מטרתנו גם במדד זה גרף זה מצטיין.
3) יכולת הפרדה: מבחינת יכולת הפרדה גם כן גרף זה מעולה בין היתר בעזרת יכולת הזום אין, אך מהיות כך שאין overlapping בין הברים השונים במיוחד מתי שנקטין את הכמות אזורים או נעשה זום אין, מדד זה גם יחשב מצוין כאן. ההפרדה הזאת תעזור למשתמש להתרכז בפיצ'ר ולא דברים שונים ולא בברים שונים. כמובן שזה רלוונטי ביותר כאשר אנו בזום אין או כמות אזורים נמוכה כי גרף בעל 175 כניסות לא יראה טוב ויסביר יותר מידי. בנוסף יש לנו את הצבעים עבור יכולת ההפרדה.
4) בצבוץ: אחת הכישלונות בבר פלוט זה חוסר היכולות ליצור בצבוץ, אך התמודדנו עם כך בעזרת צביעת כל בר לפי האזור שאליו הוא שייך כך שהוא מבצבץ וניתן להבין בצורה מעולה את ההבדלים בין האזורים בזכות זה
5) קיבוץ: השימוש באזור כקידוד לצבע עונה על קריטריון זה בצורה מעולה ועוזר לראות את ההבדלים בפיצ'ר גם בין אזורים שונים.

ועל כל הסיבות שצוינו מעל אגיד כי הגרף עונה על כל אלו ולכן הוא גרף מעולה שמתאר את מה שאני צריך ובול עונה לנו על שאלה (3).

עיבוד מקדים:

המידע נאסף מהלינק https://www.kaggle.com/datasets/unsdsn/world-happiness ובוצע עיבוד מקדים של איחוד כל הטבלאות ביחד בנוסף למידע גאוגרפי. לאחר מכן כאשר היה נראה באמצע הויזואליזציה שיש פיצ'רים אשר אין לי צורך בהם ידנית הורדתי אותם מהאקסל על מנת לא להריץ מחברת מחדש.
המחברת שבעזרתה בוצע העיבוד המקדים תוסף לzip .
יש לציין שהשמות של העמודות שונו גם דרך האקסל לעמודות שיותר התאימו לי מבחינת נוחות ואקספרסיביות למי שירצה לנבור במידע.

הסבר על הוויזואליזציה:

אצרף סרטון בתוך ה zip כיוון שלפי דעתי לוויזואליזציה יש המון פיצ'רים מאוד חזקים

כתיבת הקוד:

בכתיבת הקוד בעצם התחלתי ממעבר בגיטאהב על דוגמאות של פעולה עם streamlit וכאשר לא מצאתי משהו שמספק בסיס טוב השתמשתי בקוד של הבחור הזה https://github.com/Wilsven/World-Happiness-Index-Streamlit-App על מנת להתחיל את הפרויקט. הקוד שונה לגמרי מאז האיטרציה הראשונה לכן לדעתי אין צורך להסביר את ההבדלים כיוון שרשמתי קוד, הרבה יותר משמעותי ואינטראקטיבי וקוד זה היווה מעין בסיס לכתיבה בלבד והתחלת עבוד.
בנוסף לכך הרבה מהקוד נרשם בעזרת chatgpt ולכן יכול להיות חפיפה בין הקוד שלי לאחרים עקב כך שהוא בעצם כלי אשר אומן על דאטה קיים ונותן את אותו הפלט לאנשים שונים. אך לאורך כל הויזואליזציה השתמשתי בדברים וברעיונות שלי ועל כן אני לא מאמין שקיים עוד קוד כזה.
בנוסף לכך השתמשתי בספריות : pandas,plotly,streamlit,numpy אשר חוץ מstreamlit הינם ספריות טריוויאליות לפייתון ואין צורך לפרט.
Stremlit זה בעצם חבילה אשר מקלה על תהליך כתיבת הקוד והעיצוב ומאוד מונגשת למשתמש, אך הסיבה שבחרתי בה זה הקלות שבה ניתן לעלות את הקוד דרך חיבור לגיטאהב והעלאה למערכת שלהם על מנת להפעיל את הדאשבורד ולהשאיר אותו פועל בענן.

מקווה מאוד שנהנתם בקריאת הדוח! השקעתי בו המון וגם בקוד ויזואליזציה ובהחלט למדתי דברים חדשים על איך להנגיש מידע בצורה הרבה יותר טובה ! 😊😊😊