

פרופ' גלעד רביד



אוניברסיטת בן-גוריון בנגב  
Ben-Gurion University of the Negev

# דוח מסכם – ויזואליזציה של המידע

נושא: חברת איקומרס Olist

מגישים:

אוהד אשכנזי 318321601

מלאכי ויס 314969338

תוכן עניינים

תיאור הנתונים ואופן הטיפול .....	2
דיון בויזואליזציות קיימות .....	3
ויזואליזציה ראשונה .....	4
ויזואליזציה שניה .....	5
ויזואליזציה שלישית .....	6
ויזואליזציה רביעית .....	7
ויזואליזציה חמישית .....	8
ויזואליזציה שישית .....	9
ויזואליזציה שביעית .....	10

## תיאור הנתונים ואופן הטיפול

הנתונים בפרויקט מבוססים על מאגר נתוני איקומרס של פלטפורמת Olist וכוללים מידע רב־ממדי על הזמנות, פריטים, לקוחות, מוכרים, קטגוריות מוצרים, תשלומים ומשלוחים. בסיס הנתונים בנוי ממספר טבלאות רלציוניות המחוברות באמצעות מזהים ייחודיים (כגון מזהה הזמנה, מזהה לקוח ומזהה מוכר), ולכן שלב ראשון בתהליך כלל מיזוג (Join) בין הטבלאות לצורך יצירת טבלאות אנליטיות מאוחדות. לאחר האיחוד בוצע ניקוי נתונים שכלל טיפול בערכים חסרים, הסרת כפילויות, תיקון טיפוסים משתנים (בעיקר שדות תאריך ושעה), והמרת חותמות זמן לפורמטים אחידים המאפשרים אגרגציה לפי יום, שבוע, חודש ורבעון. בנוסף, חושבו מדדים נגזרים כגון סך הכנסות להזמנה, מחיר ממוצע לפריט, אחוז צמיחה בין תקופות, שיעור איחורים במשלוחים, מדדי ריכוזיות (Top-N), ומדדי מתאם בין משתנים מרכזיים. במקרים בהם זוהו ערכים קיצוניים חריגים (outliers), בוצע חיתוך סטטיסטי מבוקר (למשל Trim באחוזונים 1%-99%) על מנת למנוע עיוות של התמונה הכוללת. כמו כן בוצעה אגרגציה ברמות שונות – קטגוריית מוצר, תקופת זמן, אזור גיאוגרפי ומוכר – בהתאם לשאלה העסקית שנבדקה בכל ויזואליזציה. תהליך ההכנה שם דגש על עקביות מדדים, שמירה על פרשנות עסקית נכונה, והבטחת יכולת אינטראקטיבית לסינון, קיבוץ ושינוי גרנולריות בזמן אמת, כך שהנתונים יהיו לא רק נקיים ואמינים אלא גם מותאמים לניתוח חזותי תומך החלטה.



## דיון בויזואליזציות קיימות לאותם נתונים



מתוך האתר Kaggle (מאפיפה שלקחנו את הנתונים)

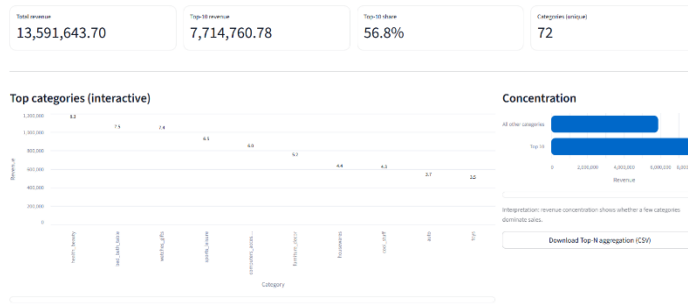
<https://www.kaggle.com/code/philipclemente/olist-e-commerce-analysis-and-dashboard/notebook>

### התמקדות בויזואליזציה השנייה והשוואה:

במסמך שלנו הצגנו ויזואליזציה מתקדמת ורב-ממדית אשר משלבת במקביל גם את נפח ההזמנות וגם את סך ההכנסות לאורך זמן, ובכך מאפשרת ניתוח דינמי של הקשר בין פעילות תפעולית לביצועים פיננסיים. זאת בניגוד לגרף הבסיסי המופיע ב-Kaggle, אשר מציג מכירות חודשיות בלבד בצורה סטטית וללא העמקה אנליטית. הויזואליזציה שלנו כוללת אינטראקטיביות מלאה – אפשרות לבצע Hover לקבלת ערכים מדויקים, לסנן טווחי זמן, לשנות רזולוציית תצוגה (שבועי/חודשי/רבעוני), ולהשוות בין המדדים. בנוסף, באמצעות מבנה של Small Multiples ניתן לבחון כל מדד בסקאלה טבעית ונפרדת, תוך שמירה על אותו ציר זמן להשוואה אינטואיטיבית. גישה זו מאפשרת לא רק לראות "כמה נמכר", אלא להבין האם העלייה במכירות מלווה בעלייה פרופורציונלית בהכנסות, האם קיימות תקופות של שחיקת ערך להזמנה, והאם הדינמיקה מצביעה על שינוי מבני בפלטפורמה – תובנות שאינן ניתנות לזיהוי בגרף בסיסי חד-ממדי.

## ויזואליזציה ראשונה

### Pareto Analysis



### ניתוח תיאורי WWH

**What** - הנתונים הם נתונים טבלאיים אגרטיביים ברמת קטגוריית מוצר. כל שורה מייצגת קטגוריה אחת, והמשתנים המרכזיים הם: סך ההכנסות (סכום מחירי הפריטים), מחיר ממוצע לפריט, מספר פריטים שנמכרו ואחוז מההכנסה הכוללת. מדובר בשילוב של משתנה קטגורי (שם קטגוריה) ומשתנים כמותיים רציפים.

**Why** - המשימה האנליטית היא זיהוי ריכוזיות ההכנסות בפלטפורמה, איתור קטגוריות דומיננטיות, והשוואה בין גודל פיננסי לבין רמת המחיר הממוצעת. בנוסף, מבקשים לבחון האם ההכנסות מפוזרות באופן שוויוני יחסית או שמא קיימת תלות במספר קטן של קטגוריות מובילות (עיקרון Pareto)

**How** - הקידוד הוויזואלי נעשה באמצעות תרשים עמודות לקידוד ההכנסות (ערוץ אורך – מדויק תפיסית), קו נוסף לקידוד המחיר הממוצע (ציר משני), ותוויות אחוזים להצגת תרומת כל קטגוריה להכנסה הכוללת. בנוסף, מוצג תרשים ריכוזיות משלים המשווה בין Top-N קטגוריות לבין שאר הקטגוריות. האינטראקטיביות מאפשרת שליטה במספר הקטגוריות המוצגות, שינוי קריטריון המיון, וסינון מחירים קיצוניים.

### הסבר בחירות העיצוב והשאלה העסקית

הגרף עונה על השאלה:

**האם ההכנסות מפוזרות על פני קטגוריות רבות, או שמספר קטן של קטגוריות מהווה את עיקר הפעילות העסקית?**

הבחירה בתרשים עמודות נובעת מהעובדה שערוץ האורך מאפשר השוואה מדויקת בין קטגוריות בדידות. הוספת קו המחיר הממוצע מאפשרת זיהוי תופעות כגון קטגוריות עם הכנסה גבוהה אך מחיר נמוך (נפח גדול), לעומת קטגוריות פרימיום עם מחיר גבוה אך נפח נמוך יותר.

הוספת אחוז מההכנסה הכוללת מעל כל עמודה מיישמת את עקרון ה־ Overview+Detail המשתמש רואה גם את הערך האבסולוטי וגם את המשמעות היחסית שלו.

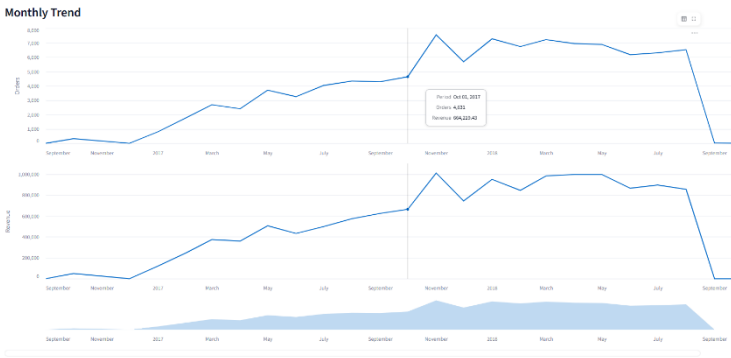
תרשים הריכוזיות המשלים (Top-N) מול היתר (מאפשר לזהות האם הפלטפורמה נשענת על מספר מצומצם של מנועי צמיחה – תובנה קריטית לניהול סיכונים עסקיים ולגיוון פורטפוליו).

היכולת לשלוט במספר הקטגוריות (Top-N) מממשת את עקרון "מבט על תחילה, ולאחר מכן זום", בכך שהיא מאפשרת למשתמש לבחון את המערכת ברמת מאקרו ולאחר מכן להתמקד בקבוצות ספציפיות.

סינון מחירים קיצוניים (trim 1%-99%) משקף עקרון סטטיסטי חשוב של הפחתת השפעת outliers לצורך קבלת תמונה יציבה ואמינה יותר.

## ויזואליזציה שניה

### Small Multiples Time Series



#### Controls

Time granularity

Monthly

☒ Show points

☐ Rolling average

☐ Log revenue (log1p)

☒ Reset zoom

### ניתוח תיאורי WWH

**What** - הנתונים הם נתוני סדרת־זמן אגרטיביים, כאשר כל רשומה מייצגת יחידת זמן (שבועית או חודשית). המשתנים הכמותיים הם: מספר הזמנות ייחודיות וסך ההכנסות הנובעות מפריטי ההזמנה. מדובר בשני משתנים רציפים בעלי סקאלות שונות, אך מבוססים על אותו ציר זמן.

**Why** - המטרה האנליטית היא זיהוי מגמות לאורך זמן, איתור תקופות האצה או האטה בפעילות, ובחינת הקשר הדינמי בין נפח פעילות (מספר הזמנות) לבין ביצוע פיננסי (הכנסות). בנוסף, מבקשים לבחון האם קיימת סינכרוניזציה בין שני המדדים או שמא קיימים פערים המעידים על שינוי במבנה הביקוש (למשל עלייה בכמות ללא עלייה תואמת בהכנסה).

**How** - הקידוד הוויזואלי מתבצע באמצעות שני תרשימי קו נפרדים (Small Multiples) החולקים ציר זמן משותף. כל מדד מוצג עם ציר Y עצמאי בעל סקאלה טבעית ומספרים מלאים.

### הסבר בחירות העיצוב והשאלה העסקית

הגרף עונה על השאלה המרכזית:

**כיצד מתפתחות ההזמנות וההכנסות לאורך זמן, והאם קיים קשר עקבי ביניהן?**

הבחירה ב־ Small Multiples במקום ציר כפול נובעת מעקרונות ויזואליזציה אקדמיים הממליצים להימנע מצירי Y כפולים כאשר קיימים פערי סדרי גודל משמעותיים. הפרדה זו מפחיתה עומס קוגניטיבי, משמרת קריאות כמותית מלאה, ומאפשרת השוואה אינטואיטיבית לאורך אותו ציר זמן.

הוספת Brush בתחתית הגרף מיישמת את עקרון: **Overview first, zoom and filter, details on demand**: המשתמש רואה את התמונה הכוללת, בוחר טווח רלוונטי, ומעמיק בניתוח מבלי לעבור דף או לשנות גרף.

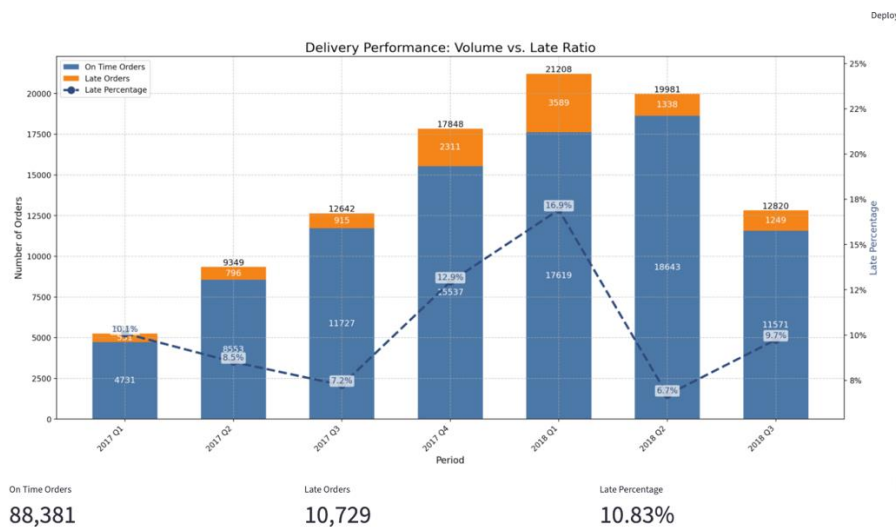
ה־ Hover מאפשר דיוק מספרי נקודתי ללא עומס טקסטואלי קבוע על הציר, וכך נשמר איזון בין מינימליזם עיצובי לבין דיוק אנליטי.

הוספת ממוצע נע מאפשרת להבדיל בין תנודתיות מקרית לבין מגמה מבנית, בהתאם לעקרונות ניתוח סדרות זמן. האפשרות לסקאלה לוגריתמית רלוונטית במצבים של צמיחה אקספוננציאלית, ומונעת דחיסה של תקופות מוקדמות.

מדדי הסיכום בראש הדף (סה"כ הזמנות, סה"כ הכנסות, וטווח תאריכים) מספקים הקשר כמותי ראשוני לפני בחינת הדינמיקה, ומיישמים עקרון: Contextual Framing.

## ויזואליזציה שלישית

### Stacked Bar Chart



Granularity

Quarterly

Date range

2017/01/01 – 2018/09/30

## ניתוח תיאורטי WWH

**What** - הנתונים הם נתונים טבלאיים הכוללים משתנה זמן כמפתח ראשי, ומשתנים של סטטוס המשלוח וכמות ההזמנות.

**Why** - המשימה המרכזית היא זיהוי מגמות עונתיות בביצועי השילוח, איתור תקופות חריגות בהן אחוז האיחורים זינק בצורה קיצונית, והשוואה בין נפח הפעילות לבין איכות הביצוע.

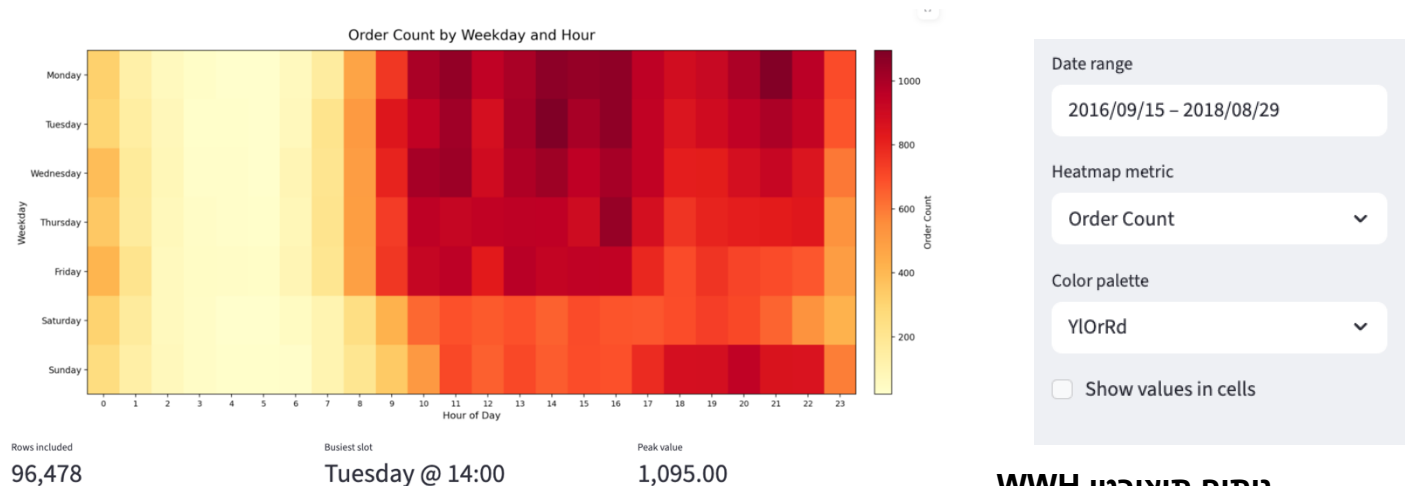
**How** - הקידוד הוויזואלי נעשה באמצעות שימוש בקווים עבור המגמה ובשטחי עמודות עבור הכמות. ערוץ האורך משמש לקידוד כמות ההזמנות מכיוון שהוא מאפשר תפיסה מדויקת יחסית, בעוד ערוץ הצבע משמש להפרדה בין הקטגוריות. בנוסף, האינטראקציה מאפשרת מניפולציה של הנתונים באמצעות סינון וקבוצת לשינוי רזולוציית הזמן.

## הסבר בחירות העיצוב והשאלה העסקית

הגרף עונה על השאלה האם הגידול בנפח ההזמנות פוגע באיכות השירות הלוגיסטי. ניתן לפלטר את רזולוציית הזמן (חודשי, רבעוני, שנתי) ואת טווח התאריכים כדי לזהות האם בעיות שילוח הן נקודתיות או כרוניות. נבחר תרשים משולב בעל ציר כפול מכיוון שהקשר בין הכמות לשיעור ההצלחה הוא קריטי. הבחירה בפלטת הצבעים נעשתה בהתאם לעקרונות הסמנטיקה של הצבע. הצבע הכחול נבחר לייצוג משלוחים בזמן מכיוון שהוא משדר יציבות ותקינות, בעוד הצבע הכתום נבחר לייצוג איחורים בשל בולטות גבוהה והיותו צבע התראה המושך את העין לבעיה. עבור קו אחוז האיחורים נבחר כחול כהה ליצירת ניגודיות גבוהה. מבחינת אינטראקציה, האפשרות לשנות את רמת הפירוט בין חודשי, רבעוני ושנתי מיישמת את עקרון "מבט על תחילה", והיכולת לסנן טווח תאריכים מאפשרת התמקדות בתקופות רלוונטיות



## ויזואליזציה רביעית Heatmap (אינטראקטיבי)



### ניתוח תיאורטי WWH

**What** - הנתונים הם רשומות של הזמנות עם חותמת זמן. שני משתנים קטגוריאליים-מחזוריים – יום בשבוע ושעה ביממה, ומשתנה כמותי מספר ההזמנות או סף ההכנסות

**Why** - המשימה היא סיכום וזיהוי דפוסים חזרתיים על פני הזמן, לצד איתור חריגים כמו שעות עומס קיצוניות

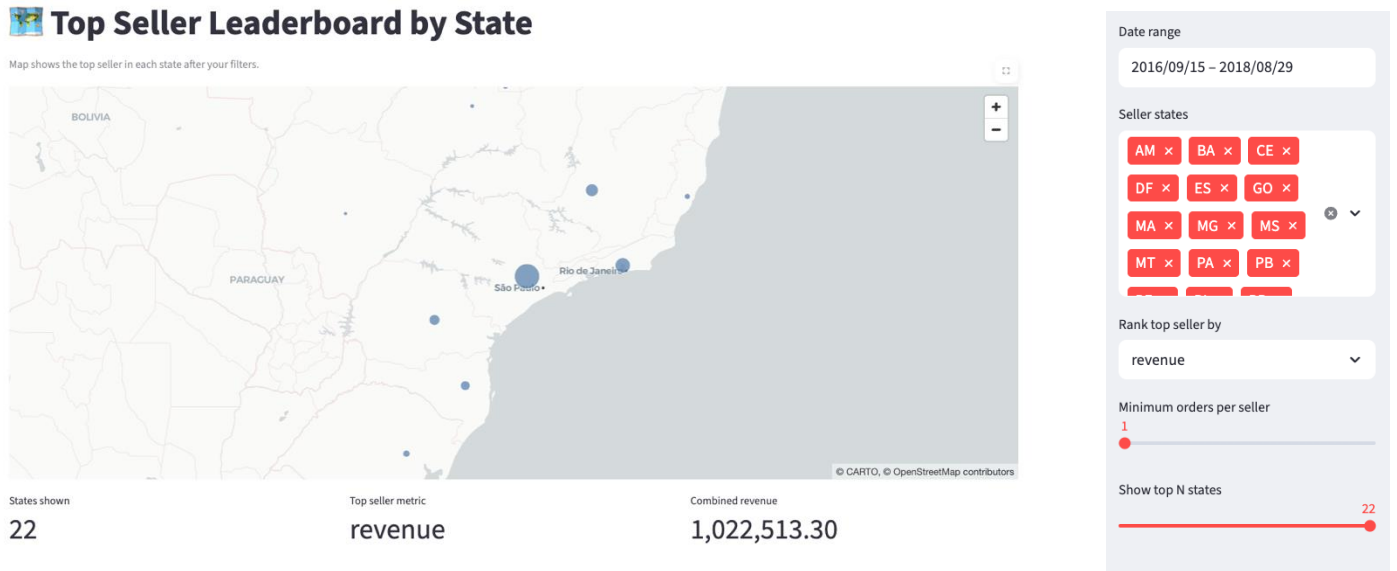
**How** - הקידוד הוויזואלי נעשה באמצעות מיקום דו-ממדי במטריצה (יום בשבוע מול שעה ביממה) ושימוש בערוץ הרוויה של הצבע לייצוג עוצמת הביקוש

### הסבר בחירות העיצוב והשאלה העסקית

השאלה העסקית שבה מתעסקת הויזואליזציה היא מתי מתרחש שיא הביקוש לצורך תכנון כוח אדם בשירות לקוחות וכדומה. בקורס למדנו שנהוג להשתמש במפות חום טמפורליות כדי לחשוף דפוסים מחזוריים של שעות וימים, כיוון שטבלאות רגילות מתקשות להציג צפיפות נתונים גבוהה כזו בצורה אינטואיטיבית. עבור זיהוי זמני השיא בחרנו במפת חום מכיוון שהיא מאפשרת סריקה מהירה של כל שעות השבוע במבט אחד ומבליטה מיידית "אזורים חמים". במפת החום נעשה שימוש בפלטות צבעים סדרתיות (כגון צהוב-כתום-אדום) המאפשרות למשתמש לתפוס באופן אינטואיטיבי את המעבר מערכים נמוכים לגבוהים.

## ויזואליזציה חמישית

### Proportional Symbol Map (אינטראקטיבי)



### ניתוח תיאורטי WWH

**What** - הנתונים הם נתונים מרחביים הכוללים מפתח גיאוגרפי דו-ממדי (קו רוחב/אורך), משתנים קטגוריאליים (מדינה, מזהה מוכר) ומשתנים כמותיים (כנסות, כמות הזמנות).

**Why** - המשימה היא איתור של מוכרים מובילים לפי אזור, והשוואה של עוצמת הפעילות העסקית בין טריטוריות גיאוגרפיות שונות

**How** - הקידוד הוויזואלי משתמש בנקודות הממוקמות לפי קואורדינטות גיאוגרפיות, תוך שימוש בערוץ הגודל (size) לייצוג המדד הכמותי הנבחר (הכנסות וכמות הזמנות).

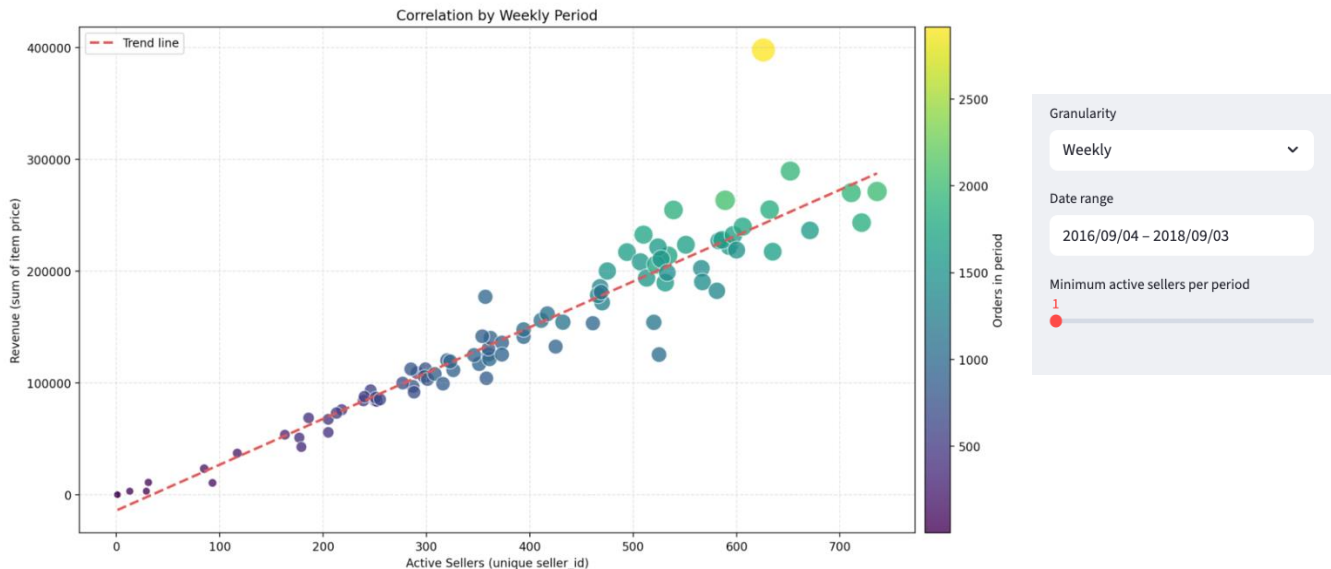
### הסבר בחירות העיצוב והשאלה העסקית

המפה עונה על השאלה "מיהם השחקנים הדומיננטיים בכל אזור גיאוגרפי ומהו היקף השפעתם". ניתן לפלטר איזורים גיאוגרפיים, טווח תאריכים, הערך הכמותי שאותו רוצים לראות על המפה (הכנסות, כמות מכירות, מספר מוכרים וכו') ומינימום הזמנה למוכר. הבחירה במפת סמלים ולא בטבלה רגילה נובעת מהחשיבות של ההקשר המרחבי – היכולת לראות האם ישנם ריכוזים של מוכרים חזקים באזורים סמוכים (למשל בדרום מזרח ברזיל) לעומת אזורים מבודדים.

הבחירה בצבע כחול שקוף עם קו מתאר לבן נועדה לשרת שתי מטרות: השקיפות מאפשרת לראות את מפת הרקע ומקלה על זיהוי חפיפות במידה וקיימות, בעוד קו המתאר הלבן מבטיח הפרדה ברורה של האובייקט מהרקע. היכולת לשנות את המדד הקובע את גודל הבועה (למשל, מעבר מהכנסות למספר הזמנות) מאפשרת למשתמש לבחון את המושג "הצלחה" מזוויות שונות.

## ויזואליזציה שישית

### Scatterplot (אינטראקטיבי)



### ניתוח תיאורטי WWH

**What** - הנתונים הם נתונים כמותיים אגרגטיביים שבהם כל פריט מייצג יחידת זמן (שבוע/חודש), והמשתנים הם כמותיים בלבד: מספר המוכרים הפעילים, סך ההכנסות ונפח ההזמנות.

**Why** - המשימה היא גילוי ואפיון של המתאם בין גודל היצע המוכרים לבין הצמיחה בהכנסות, במטרה לאשש את קיומו של "אפקט רשת" חיובי בפלטפורמה.

**How** - הקידוד הוויזואלי משתמש בנקודות בתוך תרשים פיזור, כאשר ערוצי המיקום המרחבי מייצגים את הקשר בין המשתנים, וערוצי הגודל והצבע משמשים לקידוד Redundant Encoding של נפח ההזמנות להדגשת המשקל של כל תקופה.

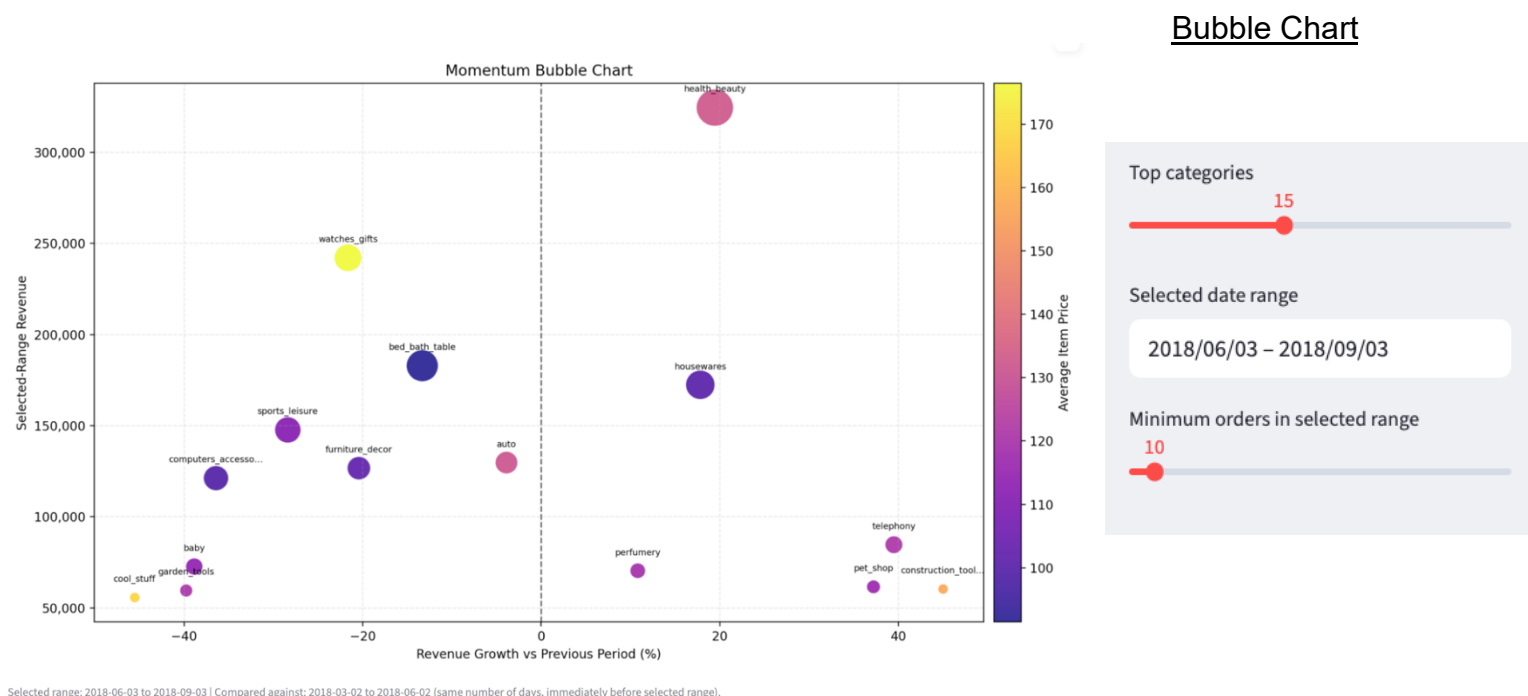
### הסבר בחירות העיצוב והשאלה העסקית

התרשים עונה על השאלה "האם קיים קשר ישיר בין הגדלת היצע המוכרים לצמיחה בהכנסות הפלטפורמה. ניתן לפלטר את רזולוציית הזמן (שבועי, חודשי, רבעוני ושנתי), טווח התאריכים, ואת מינימום המוכרים הפעילים כדי לנטרל רעשים של תקופות הקמה. הבחירה בתרשים פיזור מגיעה מהשיעורים שלמדנו על כך שגרף פיזור מתאים למציאת מתאם בין המשתנים - היכולת לראות כיצד שינוי בציר ה-X מספר מוכרים משפיע ישירות על ציר ה-Y הכנסות

הבחירה בפלטת הצבעים ובשינוי גודל הבועה לייצוג כמות ההזמנות נועדה להוסיף מימד שלישי של "נפח פעילות", המאפשר להבחין בין נקודות זמן עם הכנסות דומות אך עצימות תפעולית שונה. הוספת קו המגמה האדום המקווקו מספקת עוגן ויזואלי מיידי לכיוון ועוצמת הקשר הסטטיסטי. היכולת לשנות

את הגרנוטריות (למשל משבועי לחודשי) מאפשרת למשתמש "להחליק" תגודתיות קצרת-טווח ולהתמקד במגמות מאקרו ארוכות טווח.

## ויזואליזציה שביעית



## ניתוח תיאורטי WWH

**What** - הנתונים הם טבלה רב-ממדית שבה כל פריט מייצג קטגוריית מוצר. המשתנים כוללים משתנה קטגוריאלי (שם הקטגוריה) ומשתנים כמותיים הדורשים חישוב על בסיס חלונות זמן: אחוז הצמיחה, סך ההכנסות, נפח ההזמנות ומחיר פריט ממוצע.

**Why** - המשימה היא גילוי ואפיון של קטגוריות המכניסות ביותר בעסק, איתור קטגוריות המציגות צמיחה או דעיכה קיצונית, והשוואה בין גודל הקטגוריה לקצב השינוי שלה לצורך ניהול פורטפוליו מוצרים.

**How** - הקידוד הוויזואלי מתבצע באמצעות תרשים בועות. ערוצי המיקום משמשים למשתנים המרכזיים (צמיחה והכנסות – ציר x,y), ערוץ הגודל מקודד את נפח ההזמנות, וערוץ הצבע משמש לקידוד מחיר ממוצע למכירה של הקטגוריה.

## הסבר בחירות העיצוב והשאלה העסקית

המטריצה עונה על השאלה האסטרטגית "אילו קטגוריות הן מנועי הצמיחה העתידיים ואילו נמצאות בסיכון?". ניתן לפלטר את מספר הקטגוריות המוצגות, את טווח התאריכים להשוואה ואת מינימום ההזמנות כדי לסנן קטגוריות נישה לא יציבות. הבחירה בתרשים בועות, המהווה הרחבה של תרשים פיזור, נובעת מהצורך להציג ארבעה ממדים של מידע בו-זמנית כדי לקבל תמונה הוליסטית של "בריאות הקטגוריה": האם היא גדולה  $y$ , האם היא צומחת  $x$ , האם היא פופולרית (גודל), והאם היא יקרה (צבע).

הוספת קו הייחוס האנכי בערך 0 (צמיחה אפסית) היא קריטית לחלוקת המרחב לאזורים סמנטיים ברורים: צד ימין מייצג הזדמנות (צמיחה חיובית) וצד שמאל מייצג סיכון (דעיכה). הבחירה בפלטת הצבעים עבור מחיר קטגוריה הממוצע מאפשרת לזהות האם הצמיחה בקטגוריה מסוימת מונעת על ידי מוצרים זולים ונגישים או על ידי מוצרי פרימיום יקרים, תובנה שהיא חיונית לתכנון מלאי ומבצעים.