**Homework 1 – Business Intelligence**

| **שם** | **ת.ז** |
| --- | --- |
| תומר בן יהודה | 307854422 |
| אלעד חזיזה | 312326788 |
| עידן כהן | 313522534 |

**חלק 1**

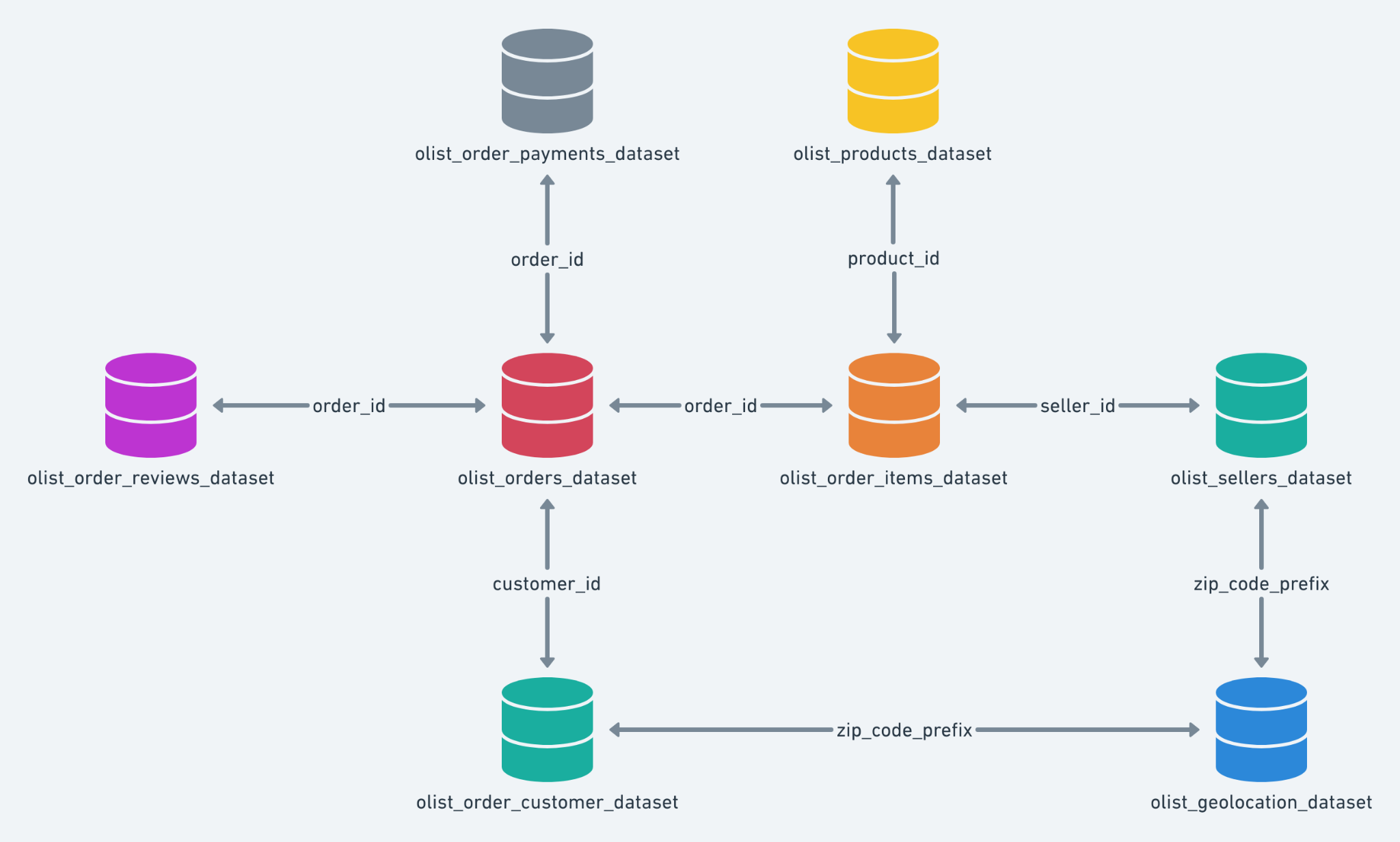
בחרנו את אוסף הנתונים הבא:

**Brazilian E-Commerce Public Dataset by Olist**

<https://www.kaggle.com/olistbr/brazilian-ecommerce>

מבנה הנתונים מכיל 100K מכירות שבוצעו בין השנים 2016-2018 בחנות Olist בברזיל.

בחרנו במבנה נתונים זה משום שהוא מקיף ומאפשר לעבוד עם מספר טבלאות במקביל:



**חלק 2**

1. אלו שאלות המחקר שבחרנו לאוסף הנתונים שנבחר:

**Supervised**

* מה הוא פוטנציאל ההכנסות של חנות ברשת?

**Unsupervised**

* לאילו קבוצות ניתן לחלק את הלקוחות של החנות? (לטובת קמפיין פרסומי)

2.אלו הKPIs לשאלות מחקר:

א. שאלת המחקר היא: מה הוא פוטנציאל הרווחיות של חנות ברשת?

| **מדד** | **הסבר SMART** |
| --- | --- |
| **סה"כ מכירות במסגרת זמן** | ברור מה המודדים,בכל מכירה יש כסף שניתן למדוד אותו,נתוני המכירות נאספים בכל מכירה,מטרת הארגון היא למכור לכן הוא רלוונטי ויש הגבלה ברורה בזמן |
| **גודל מכירה ממוצע** | קל להבין מה מודדים, קיים בנתונים מספר המכירות וכמות הכסף, נתוני המכירות נאספים בכל מכירה ונותן קנה מידה למידת הצלחת הארגון |
| **מספר מכירות לחנות** | מדד ספציפי לחנות, מדיד בגלל הצורך לשלם בסוף עסקה, ניתן להשיג את הנתונים בקלות רלוונטי מאוד למטרת הארגון |
| **רווח נקי לחנות** | ברור מה מודדים, ניתן למדוד אותו בקלות והוא המדד החשוב והרלוונטי ביותר לעסק |
| **מספר הכנסות ממוצע ללקוח בחנות** | המדד הוא SMART כי הוא רלוונטי, עונה על צורך חשוב לעסק, קל למדידה והוא עצם הקריאה שלשהשם שלו מעבירה מה הוא בודק |

ב. שאלת המחקר היא:לאילו קבוצות ניתן לחלק את הלקוחות של החנות?

| **מדד** | **הסבר SMART** |
| --- | --- |
| **כמה לקוחות מגיעים מכל עיר** | המדד מסביר את עצמו בשם, רלוונטי לשאלת המחקר משום שחלוקה גיאוגרפית יכולה להיות גם חלוקה לקבוצות וניתן למדידה משום שיש כתובת משלוח לכל הזמנה |
| **גודל מכירה ממוצע** | קל להבין מה מודדים, קיים בנתונים מספר המכירות וכמות הכסף, נתוני המכירות נאספים בכל מכירה ונותן קנה מידה למידת הצלחת הארגון |
| **מספר מכירות בחודש X לעיר Y** | המדד בודק עונתיות. הוא ניתן למדידה בקלות, הוא רלוונטי משום שכך ניתן להבין מתי כדי להגביר את הפרסום לקבוצה מסוימת |
| **ממוצע מספר רכישות ללקוח** | בעזרת מדד זה ניתן למצוא כמה רכישות ללקוח ולשייך אותו לקבוצה. ניתן למדוד אותו בקלות משום שיש מזהה לכל לקוח של אמצעי תשלום, והוא רלוונטי לשאלה משום שכך ניתן לאפיין התנהגויות של קבוצות שונות. |
| **מהי תדירות הרכישה המומצעת ללקוח** | מדד זה בודק את התדירות הקנייה של לקוח. הוא SMART משום שהוא רלוונטי לאפיון דפוס התנהגות, פשוט למדידה ועונה על צורך חשוב לעסק |

**חלק 3**

את כל הניתוחים ביצענו בעזרת מחברת notebook באתר של kaggle. להלן הקישור:  
<https://www.kaggle.com/eladhaziza/bi-project>

1. מדדי פיזור של התכונות:  
   תחת הכותרת "Statistical Information" בדף הקאגל המצורף (מחולק לעמודות שמיות ונומריות)
2. קורלציה:  
   תחת הכותרת "Correlation - Numerical Values" בדף הקאגל המצורף, נוסף על כך, ביצענו ניתוח קשר בין עמודות שמיות מעניינות לנומריות תחת הכותרת "Statistical Information".
3. מדדים סטטיסטים נוספים:  
   תחת הכותרת "Statistical Information" בדף הקאגל המצורף.

ניתן להסיק מן הנתונים כי ב- Bahia ממוצע הרכישה של מוצר לכל עסקה הוא הגבוה ביותר, מניתוח נוסף אנו יכולים לקשור עובדה זו למעמד הסוציו-אקונומי באזור זה.

בנוסף, ניתן לראות כי כ-74% מכלל העסקאות במאגר המידע מבוססות על שימוש בכרטיס האשראי וזאת ללא קשר למספר עסקאות חריגות שבוצעו בכרטיסי האשראי במאגר המידע.