BI SPECIALIST CHALLENGE

BY NICK STOCKEN

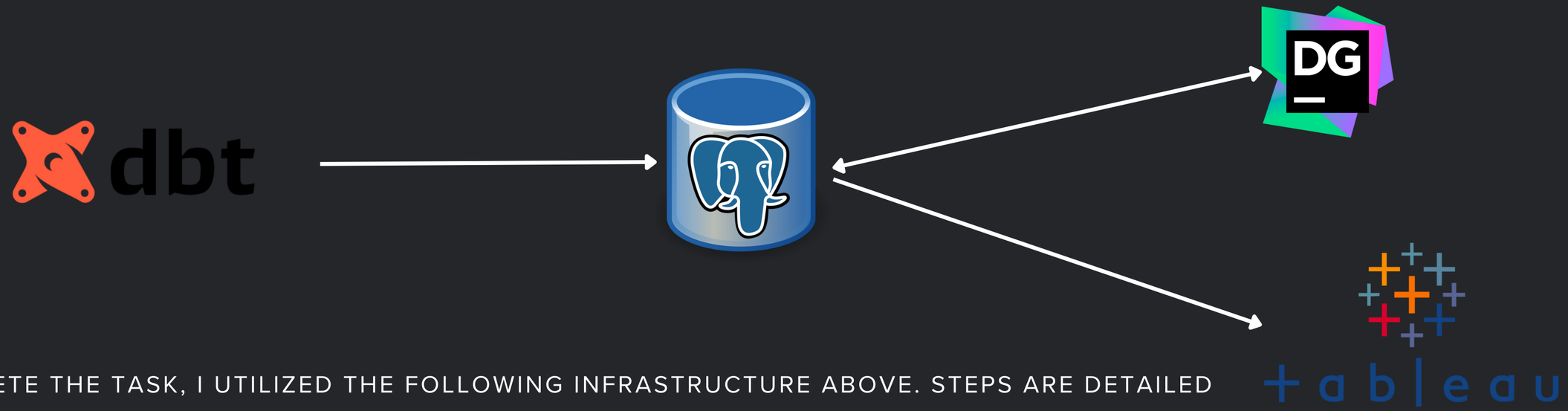
TABLE OF CONTENTS

REPOSITORY CONTENTS

https://github.com/nstocken/iconic

- ALL DATA SITS WITHIN THE DBT PROJECT FOLDER “ICONIC-PROJECT”
- BOTH THE STAGING(CLEANING) AND FINAL MODELS FOR THE DATA TRANSFORMATION QUESTION SECTION EXIST WITHIN THE “MODELS” FOLDER.
- ALL THE QUERIES FOR THE DATA ANALYSIS SECTION SIT IN THE “OTHER_QUERIES” FOLDER.
- ALL TABLEAU RELATED FILES SIT IN THE TABLEAU FOLDER INCUDING A PDF VERSION OF THE DASHBOARD.
- ALL OTHER FOLDERS ARE REDUNDANT AS THEY ARE SIMPLY PART OF THE DBT PROJECT

TECH SETUP



TO COMPLETE THE TASK, I UTILIZED THE FOLLOWING INFRASTRUCTURE ABOVE. STEPS ARE DETAILED BELOW:

1. SET UP A LOCAL POSTGRESQL SERVER.
2. INSTALLED DBT CORE FOR POSTGRESQL AND CREATED A NEW PROJECT.
3. CONNECTED DBT AND DATAGRIP TO THE POSTGRESQL DATABASE.
4. UPLOADED CSV DATA VIA DATAGRIP TO INITIATE MODELING IN DBT.
5. UTILIZED DBT FOR DATA MODELING.
6. DOWNLOADED TABLEAU AND LINKED IT TO THE POSTGRESQL DATABASE FOR CREATING VISUALIZATIONS.

DATA QUALITY AND ISSUES

DATASET	ISSUES	FIXES
USERS	• ADDRESSES COLUMN INCLUDES OTHER KEYS OTHER VALUES OTHER THAN ADDRESS I.E FIRST_NAME, LAST_NAME. • JSON USES DOUBLE QUOTES FOR KEY NAMES AND STRING VALUES • BOOLEAN VALUES SHOULD BE REPRESENTED AS TRUE OR FALSE IN JSON (ALL LOWERCASE)	• I IGNORE THESE FIELDS IN THE ANALYSIS AND ONLY EXTRACT THE INFORMATION REQUIRED. • I REPLACE THE SINGLE QUOTES WITH DOUBLE QUOTES • I LOWER CASE THE BOOLEAN VALUES
PRODUCTS	• THE STOCK AND PRICE COLUMN HAS SOME NEGATIVE VALUES. AS STATED IN THE TASK DESCRIPTION "STOCK IS ALWAYS POSITVE".	• I MADE THESE VALUES POSITIVE BY TAKING THE ABSOLUTE VALUE. HOWEVER, IN A REAL SITUATION THESE SHOULD BE FURTHER INVESTIGATED TO AVOID FALSE STOCK VALUES.
INTERACTIONS	• THE TIMESTAMP IS A UNIX TIMESTAMP. THIS SHOULD BE CONVERTED ESPECIALLY FOR READABILITY AND REPORTING PURPOSES. • USER_ID IS NULL MANY TIMES. AS STATED IN THE TASK DESCRIPTION THIS COLUMN SHOULD BE NOT NULL. • FOR THE SAME USER_ID AND ITEM_ID COMBINATION WE SEE THE SAME EVENT FIRE AT THE EXACT SAME TIME(MAINLY PRODUCT VIEWED)	• CONVERTED TO TIMESTAMP YYYY-MM-DD HH:MM:SS.SSSSSS • REMOVED ALL EVENT DATA WHERE THE USER ID IS NULL • THIS DUPLICATION IS REMOVED.

OTHER ASSUMPTIONS

- WHEN SESSIONIZING THE ACTIVITY DATA, WE EMPLOY A 30-MINUTE ACTIVITY WINDOW FOR EACH USER_ID/ITEM_ID COMBINATION, IF A SUBSEQUENT EVENT OCCURS AFTER 30 MINUTES, WE INITIATE A NEW SESSION. WITHIN A SESSION. IF THE SAME EVENT, PARTICULARLY THE PRODUCT VIEWED EVENT OCCURS TWICE, WE CONSIDER THE TIMESTAMP OF THE MAXIMUM OCCURRENCE.

- IF A DISCOUNT IS APPLIED AT ANY STAGE DURING THE TRANSACTION JOURNEY, THE DISCOUNT HOLDS TRUE FOR THAT ENTIRE JOURNEY.

- ORDER COMPLETED FOR 2 OR MORE ITEM IDS NEVER OCCURS AT THE SAME TIME. CONSEQUENTLY, THE MODEL ASSUMES A ONE-PRODUCT TYPE PER ORDER SCENARIO.

- THE PRODUCT ADDED EVENT NEVER TRIGGERS MULTIPLE TIMES WITHIN THE SAME SESSION. AS A RESULT, THE MODEL ASSUMES THAT AN ORDER CONTAINS A QUANTITY OF 1 FOR THE PRODUCT.

- ANOTHER OPTION IS TO SOLELY PARTITION BASED ON USER_ID, CONSIDERING THE POSSIBILITY THAT EVENTS MIGHT OCCUR ON THE CURRENT ITEM_ID BUT RELATE TO PREVIOUS ITEMS IN THE SESSION. HOWEVER, THIS METHOD APPEARED UNCONVENTIONAL GIVEN THE STRUCTURE OF THE INTERACTION DATA. I OPERATE UNDER THE ASSUMPTION THAT AN EVENT OCCURS FOR EVERY ITEM INVOLVED IN THE TRANSACTION. FOR EXAMPLE. IF AN ORDER COMPLETED WITH ITEMS IN THE CART, EVERY ITEMS ORDER COMPLETED TIMESTAMP WOULD BE AT THE SAME TIME.

ANALYSIS QUESTIONS

FOR THIS SECTION PLEASE REFER TO THE QUERIES FOLDER FOR THE REQUIRED SQL.

1) WHICH EVENT HAS LOW TRANSITION RATE AND CAN YOU LET US KNOW THE TRANSITION RATE ACROSS EACH OF THE EVENTS?

SEE BELOW TRANSISTION RATES. PRODUCTVIEWED --> PRODUCTADDED HAS THE LOWEST RATE.

- PRODUCTVIEWED --> PRODUCTADDED : 32%
- PRODUCTADDED --> CARTVIEWED: 67%
- CARTVIEWED --> CHECKOUTSTARTED: 44%
- CHECKOUTSTARTED --> ORDERCOMPELTED: 51%

2) WHAT IS THE PERCENTAGE OF CART ABANDONMENT ACROSS THE STORE?

- 49%

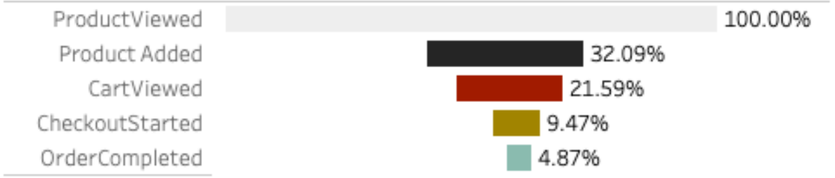
3) FIND THE AVERAGE DURATION BETWEEN CHECKOUT STARTED AND ORDER COMPLETED AND DO YOU FIND ANY ANOMALY IN THE DATA?

- 2.5 SECONDS
- THE AVERAGE TRANSACTION SPEED IS ABNORMALLY FAST. AN ANOMALY SPOTTED IS THAT SOME OF THE CHECKOUT STARTED EVENTS OCCUR THE EXACT SAME TIME AS THE ORDER COMPLETE EVENTS.

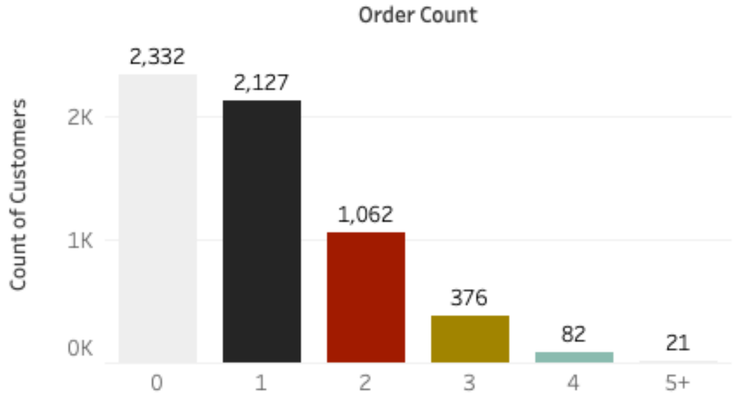
THE ICONIC

KEY DATA INSIGHTS

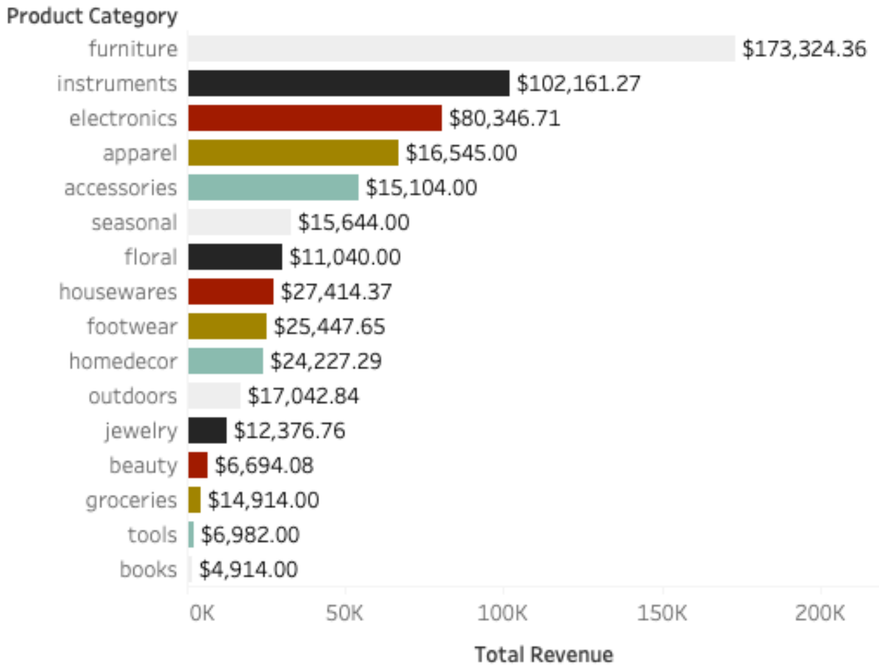
Overall Funnel



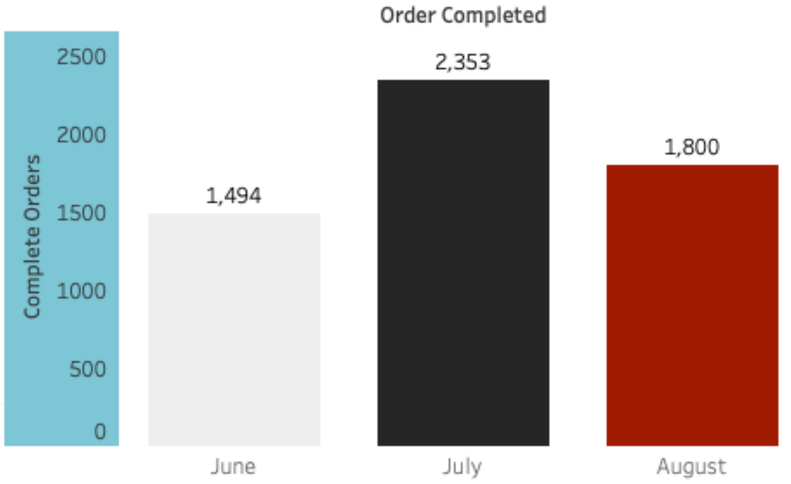
Lifetime Orders Distribution



Total Revenue by Category



Orders Completed per Month



KEY INSIGHTS

- ONLY A FRACTION (SPECIFICALLY 4.87%), OF USERS WHO INITIALLY VIEW A PRODUCT END UP COMPLETING AN ORDER, WHICH TRANSLATES TO JUST 1 IN 20 PEOPLE. THIS DISCREPANCY SUGGESTS THAT A SIGNIFICANT NUMBER OF USERS ARE IN THE BROWSING PHASE GIVEN THE SUBSTANTIAL DROP-OFF FROM PRODUCT VIEWS TO THE PRODUCT ADDED STEP.

- INTERESTINGLY, THE TWO HIGHEST-EARNING CATEGORIES(NAMELY FURNITURE AND INSTRUMENTS) CONTRIBUTE SIGNIFICANTLY TO THE TOTAL REVENUE, ACCOUNTING FOR 42% OF THE OVERALL REVENUE.

- UPON ANALYZING THE THREE MONTHS OF PROVIDED DATA, IT'S NOTABLE THAT JULY STANDS OUT WITH THE HIGHEST NUMBER OF COMPLETED ORDERS. THIS SPIKE IN ACTIVITY COULD POTENTIALLY BE ATTRIBUTED TO END-OF-FINANCIAL-YEAR SALES, INFLUENCING USER BEHAVIOR.

- EXAMINING THE CUSTOMER BASE OVER THE THREE-MONTH PERIOD REVEALS THAT 41% OF CUSTOMERS HAVE NOT MADE ANY TRANSACTIONS, WHILE 38% HAVE MADE ONLY ONE ORDER. THIS PATTERN SUGGESTS THAT A CONSIDERABLE PORTION OF CUSTOMERS ENGAGE INFREQUENTLY IN TRANSACTIONS, HIGHLIGHTING POTENTIAL OPPORTUNITIES FOR TARGETED ENGAGEMENT AND RETENTION STRATEGIES.